Large time step TVD IMEX Runge-Kutta schemes based on arbitrarily high order Butcher tableaux

Victor Michel-Dansac, Andrea Thomann

Large time step TVD IMEX Runge-Kutta schemes based on arbitrarily high order Butcher tableaux

Victor Michel-Dansac*, Andrea Thomann†

March 25, 2021

Abstract

The context of this work is the development of large time step TVD IMEX Runge-Kutta schemes to approximate the solution of hyperbolic multi-scale equations. A key feature of our newly proposed TVD schemes is that the resulting CFL condition does not depend on the large-scale eigenvalues of the multi-scale PDE, as long as they are treated implicitly. However, a result from Gottlieb et al. [15] states that unconditionally stable implicit TVD or L^∞ stable RK schemes can only be of first order. We show that this result is also valid for IMEX-RK schemes, which have a CFL restriction only depending on the explicitly treated scales. Therefore, the goal of this work is to improve the precision of a first-order IMEX-RK scheme, while retaining its L^∞ stability and TVD properties. In this work we extend and generalize the approach introduced in [9] based on a convex combination between a first-order TVD IMEX Euler scheme and a potentially oscillatory high-order IMEX-RK scheme. We derive and analyse the method for a scalar multi-scale equation and we numerically assess the performance of our TVD schemes compared to standard L-stable and SSP IMEX RK schemes from the literature. Finally, we combine our TVD schemes with the MOOD framework to increase their applicability. For numerical validation, we apply the schemes to the isentropic Euler equations and compare the results with [9] where a second order scheme is used as a basis of the TVD scheme.

1 Introduction

Multi-scale equations arise in a wide range of applications, such as shallow water equations studied e.g. in [3], magnetohydrodynamics [25], multi-material [**?**] or atmospheric flows [23]. When developing numerical methods for such applications, it is of prime importance to obtain physically admissible solutions under these multi-scale constraints.

In order to numerically treat these different scales, one must assess whether the fast scales are relevant to the physical solution. Indeed, accurately capturing these fast scales requires a very restrictive time step. This issue is discussed e.g. in [16] for the Euler equations. When the impact of the fast scales on the physical solution is less important, numerical methods which do not accurately capture all scales but follow only the slow dynamics are necessary. One option, which we will study in this paper, is to use Implicit-Explicit (IMEX) schemes, where the terms associated to the fast wave propagation are treated implicitly. Those schemes are well studied in the literature, see for instance [2] for efficient IMEX schemes applied on hyperbolic-parabolic problems, [31] for IMEX schemes adapted to stiff relaxation source terms, or [29, 10, 4] for IMEX schemes designed for the

*Université de Strasbourg, CNRS, Inria, IRMA, F-67000 Strasbourg, France; victor.michel-dansac@inria.fr
†Institut für Mathematik, Johannes Gutenberg-Universität Mainz, Germany; athomann@uni-mainz.de

low Mach regime of the Euler equations, as well as the references given therein. Therefore, in this work, we are concerned with hyperbolic systems whose stiffness comes from the flux, rather than a source term. Let us emphasise that we will not consider hyperbolic systems with stiff source terms typically arising from relaxation processes. For their treatment, we refer for instance to [31].

Higher order schemes are known to introduce spurious oscillations in the solution away from smooth regions. This is an issue, especially when considering non-linear hyperbolic equations, as the solution can develop discontinuities even when starting with a smooth initial condition. This was already observed by Harten in [17], who introduced the notion of total variation diminishing (TVD) schemes, and constructed non-oscillatory explicit and implicit second-order TVD schemes. Those schemes are non-linear, even when applied on linear equations, as from Godunov's theorem follows that linear TVD schemes can only be of first-order [11]. Since non-linear implicit schemes are very computationally costly, especially when applied to non-linear systems of equations, the construction of higher order explicit TVD schemes remained an active area of research, see e.g. [34, 36, 14] and references therein. Later, in the more general framework of strong stability preserving (SSP) implicit and explicit schemes [15], the stability property is achieved by relying on convexity arguments regarding forward and backward Euler schemes, rather than adding artificial viscosity to achieve the TVD property, as was done in [17, 36, 32]. The high-order explicit and implicit SSP schemes developed in [13, 12, 15] have a CFL restriction of the order of the CFL restriction of a forward Euler scheme. This makes the use of high-order implicit SSP schemes rather costly and impractical in applications compared to high-order explicit SSP schemes, as was remarked in [13]. Regarding IMEX SSP schemes, we refer to [18, 7, 19]. All high-order SSP schemes mentioned above require the time step to depend on all scales to achieve stability, but are provably of high order. Unfortunately, they are not well suited for the multi-scale setting, where the time step is strongly restricted by the fast scale leading, in extreme cases, to a vanishing time step.

In contrast, our focus here is the construction of large time step IMEX TVD schemes, which means that the CFL restriction solely stems from the explicitly treated terms. The work presented in this manuscript is greatly motivated by the seminal work by Gottlieb et al. [15], where it was proven that an unconditionally TVD implicit RK scheme is at most first-order accurate. Unfortunately, this result holds also for IMEX discretisations with a scale-independent CFL restriction, whose proof we have included for completeness in Appendix A. In fact, this discouraging result is also observed in [9, 5] when attempting to construct second-order TVD IMEX schemes for the Euler equations.

In the present work, we seek in a first step the design of first-order TVD IMEX RK schemes that have a higher resolution than the standard first-order backward/forward Euler IMEX scheme. The approach given here builds on the results from [9, 28], where the increase in precision is achieved by introducing a convex combination of said first-order TVD scheme with an oscillatory second-order scheme. In [9], the ARS(2,2,2) scheme from [2] is used as a basis for the convex combination, and this result was extended to a general class of second-order IMEX RK schemes in [28]. Here, we generalize and extend the results from [9, 28] further, to a convex combination with arbitrarily high order schemes. Note that convex combinations have already been used to recover first-order properties lost at higher orders, see for instance [20] to recover the positivity property or [27] for well-balanced problems.

As the TVD property is crucial to accurately capture discontinuities in the numerical solution, it is of less importance in smooth regions. In order to achieve a high-order approximation of the solution in such regions, while keeping the solution oscillation-free in the vicinity of discontinuities, we adapt a MOOD-like procedure, introduced in [6], to the case of IMEX schemes. In this framework, our first-order TVD schemes can be used as a correction when the solution computed with a high-order IMEX scheme of the reader's choice leaves the physical admissibility domain. This makes our TVD-IMEX-MOOD schemes interesting for applications, as a higher order approximation in smooth regions and an oscillation-free shock description can be achieved.

The paper is organised as follows. In section 2, we describe the problem of multi-scale equations, illustrated by a scalar linear hyperbolic equation. We shortly recall the IMEX formalism to numerically approximate those stiff equations, and prove the order restriction for the construction of high-order TVD IMEX-RK schemes with a scale independent time step restriction. In section 3, we derive a TVD IMEX scheme based on a convex combination between a second-order and a first-order IMEX update. The problem of a TVD space discretisation is also addressed in this section. The extension to TVD schemes based on arbitrarily high order Butcher tableaux is discussed in section 4. Therein, we show that the convex combination on the time updates of the the first- and high-order IMEX schemes is not enough to find a TVD scheme. Instead, we also apply a convex combination at each stage of the IMEX scheme. This novel method is illustrated by the construction of a TVD scheme based on third-order tableaux, combined with a third-order limiting procedure for the explicit space discretisation. Section 5 is devoted to numerical experiments to verify the necessity of large time step TVD IMEX RK methods for multi-scale problems. First, we introduce a MOOD procedure adapted to our TVD IMEX schemes. We give a strategy on how to find optimal values for the free parameters of the underlying TVD schemes, by compromising between precision and CPU time. To numerically validate that our TVD-IMEX-MOOD schemes are a noticeable improvement over widely used L-stable IMEX and the SSP IMEX schemes, we compare the performance of the schemes in terms of accuracy, CPU times and CFL restrictions on continuous and discontinuous solutions of the scalar multi-scale equation. We finally apply the scheme to the isentropic Euler equations. To complete this manuscript, a conclusion is presented in section 6.

2 Problem description

We consider the scalar linear two-scale initial value problem

$$\begin{cases} w_t + c_m w_x + \dfrac{c_a}{\varepsilon} w_x = 0, \\ w(0,x) = w^0(x), \end{cases} \tag{2.1}$$

where $w : (\mathbb{R}^+, \Omega) \to \mathbb{R}$, $\Omega \subset \mathbb{R}$. In (2.1), c_m and c_a/ε respectively represent a slow and a fast transport velocity. Note that the fast velocity can be adjusted by different choices of $\varepsilon > 0$. Thereby, c_m and c_a are considered independent of ε. Without loss of generality, we consider only the positive transport direction, i.e. $c_m, c_a > 0$.

The toy model (2.1) mimics, in a simplified linear manner, the wave structure of e.g. the Euler equations, see for instance [9]. In this case, the system is characterised by a large pressure gradient in the low Mach number regime, resulting in fast acoustic wave speeds. These fast speeds are represented by c_a/ε in our toy model, where ε acts as the Mach number squared. The Mach number independent advection speeds are described by the velocity c_m in our toy model (2.1).

Nevertheless, when developing numerical methods for the simplified scalar case (2.1) with small $\varepsilon > 0$, one faces similar challenges as for hyperbolic equations with large gradients in the flux function. Treating both derivatives in (2.1) explicitly leads to the following CFL condition, which depends on ε to ensure stability:

$$\Delta t \leqslant \varepsilon \, \mathcal{C} \, \frac{\Delta x}{\varepsilon \, c_m + c_a},$$

where $\mathcal{C}$ is a CFL coefficient independent of ε. Thus, when ε tends to zero, the time step Δt tends to zero as well. Using an explicit scheme in this regime becomes very costly in terms of computational time. Therefore, we adopt an IMEX approach and treat the derivative associated with the fast speed c_a/ε implicitly, whereas the one associated with the slow speed c_m remains explicit, to yield the following time step restriction independent of ε:

$$\Delta t \leqslant \widetilde{\mathcal{C}} \, \frac{\Delta x}{c_m}.$$

Since our main goal is to derive an L^∞ stable and TVD scheme, we have to use an upwind discretisation for both derivatives. This is motivated by the results in [10], where it is shown that for a non-linear system centred differences destroy the L^∞ stability. Although our setting is linear, we avoid centred differences to be able to apply the approach developed here on non-linear systems such as the isentropic Euler equations discussed in Section 5.4.

The space and time discretisation follows the usual finite difference framework, although it can be easily translated into the finite volume setting. The space domain Ω is partitioned in N uniformly spaced points $(x_j)_{j\in\{1,\dots,N\}}$ with the step size Δx. We discretise the time variable with $t^n = n\Delta t$, where Δt denotes the time step. Then the solution $w(t,x)$ at (t^n, x_j) is approximated by w_j^n. A semi-discrete first-order approximation of (2.1) in space, with $\Delta_j(t) = w_j(t) - w_{j-1}(t)$, is given by

$$\partial_t w_j(t) + \frac{c_m}{\Delta x}\Delta_j(t) + \frac{c_a}{\varepsilon\Delta x}\Delta_j(t) = 0. \tag{2.2}$$

Later on, to extend the space discretisation in (2.2) to higher orders, we will use a high-order reconstruction combined with a limiting procedure to ensure the TVD property.

We first discuss the time integration in section 2.1. Then, we introduce the technique that we will use to derive more precise TVD first-order schemes in section 2.2.

2.1 High-order IMEX Runge Kutta time integration

For the time integration of (2.2), we use the IMEX-RK framework. The time update for an s-stage IMEX-RK scheme for equation (2.2) is given by

$$w_j^{n+1} = w_j^n - \lambda\sum_{k=1}^{s}\tilde{b}_k\Delta_j^{(k)} - \mu_\varepsilon\sum_{k=1}^{s} b_k\Delta_j^{(k)}, \tag{2.3}$$

where we have set

$$\lambda = \frac{\Delta t}{\Delta x}c_m, \qquad \mu_\varepsilon = \frac{\Delta t}{\Delta x}\frac{c_a}{\varepsilon}, \qquad \Delta_j^{(k)} = w_j^{(k)} - w_{j-1}^{(k)},$$

and where the stages are defined as

$$w_j^{(k)} = w_j^n - \lambda\sum_{l=1}^{k-1}\tilde{a}_{kl}\Delta_j^{(l)} - \mu_\varepsilon\sum_{l=1}^{k} a_{kl}\Delta_j^{(l)}. \tag{2.4}$$

The weights $\tilde{a}_{kl}$, a_{kl} appearing in the definition (2.4) of the stages $w^{(k)}$, and $\tilde{b}_k$, b_k in the update w^{n+1} given by (2.3), are summarized in two triplets $(\tilde{A},\tilde{b},\tilde{c})$ and (A,b,c), with $\tilde{A}, A \in \mathbb{R}^{s\times s}$, $\tilde{b}, b \in \mathbb{R}^s$. The coefficients $\tilde{c}, c \in \mathbb{R}^s$ contain the intermediate time steps associated to the respective computational stages. Here, we consider the matrix associated to the explicit part $\tilde{A}$ to be lower triangular with zeros on the diagonal, and the matrix connected to the implicit part A to be lower triangular, resulting into a DIRK (diagonally implicit RK) scheme. Since we are considering multi-scale equations, we wish, for computational efficiency, for the CFL restriction of the resulting scheme to only depend on the slow scale associated with λ. In addition, for the sake of illustrating our approach, we consider an IMEX-RK method of type CK (Carpenter and Kennedy) [21], i.e. we take the first row of A to be zero. This choice, as it was shown in detail in [28] for a generic second-order CK method, requires the first column of A to be zero, as well as b_1, to ensure a CFL condition independent of ε. We give

the structure in the following Butcher tableaux notation:

$$\text{explicit:}\quad \begin{array}{c|cccc} 0 & 0 & 0 & \cdots & 0 \\ \tilde{c}_2 & \tilde{a}_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \tilde{c}_s & \tilde{a}_{s1} & \cdots & \tilde{a}_{s,s-1} & 0 \\ \hline & \tilde{b}_1 & \cdots & \tilde{b}_{s-1} & \tilde{b}_s \end{array} \qquad \text{implicit:}\quad \begin{array}{c|cccc} 0 & 0 & 0 & \cdots & 0 \\ c_2 & 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & 0 & a_{s2} & \cdots & a_{ss} \\ \hline & 0 & b_2 & \cdots & b_s \end{array}, \tag{2.5}$$

where the coefficients $\tilde{c}$ and c are respectively connected to $\tilde{A}$ and A via

$$\tilde{c}_i = \sum_{j=1}^{i-1} \tilde{a}_{ij} \quad \text{and} \quad c_i = \sum_{j=1}^{i} a_{ij}. \tag{2.6}$$

For an approach based on a different structure than (2.5), where the first column of A is nonzero, see Appendix C.

We are interested in higher order Butcher tableaux, with an order $p \geqslant 1$. This implies that the weights have to fulfil high-order compatibility conditions. The order conditions to obtain a scheme up to order three are given in Table 1 taken from [30]. For orders higher than three, we refer to the order conditions in [21].

Table 1: Order conditions for IMEX-RK schemes up to third-order

First-order:	$\sum_{k=1}^{s} \tilde{b}_k = 1,$	$\sum_{k=1}^{s} b_k = 1$		
Second-order:	$\sum_{k=1}^{s} \tilde{b}_k \tilde{c}_k = \frac{1}{2},$	$\sum_{k=1}^{s} b_k c_k = \frac{1}{2},$	$\sum_{k=1}^{s} \tilde{b}_k c_k = \frac{1}{2},$	$\sum_{k=1}^{s} b_k \tilde{c}_k = \frac{1}{2}$
Third-order:	$\sum_{k=1}^{s} \tilde{b}_k \tilde{c}_k^2 = \frac{1}{3},$	$\sum_{k=1}^{s} b_k c_k^2 = \frac{1}{3},$	$\sum_{k=1}^{s} \tilde{b}_k \tilde{c}_k c_k = \frac{1}{3},$	$\sum_{k=1}^{s} b_k \tilde{c}_k c_k = \frac{1}{3},$
	$\sum_{k,l=1}^{s} \tilde{b}_k \tilde{a}_{kl} \tilde{c}_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} \tilde{b}_k \tilde{a}_{kl} c_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} \tilde{b}_k a_{kl} \tilde{c}_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} \tilde{b}_k a_{kl} c_k = \frac{1}{6},$
	$\sum_{k,l=1}^{s} b_k \tilde{a}_{kl} \tilde{c}_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} b_k \tilde{a}_{kl} c_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} b_k a_{kl} \tilde{c}_k = \frac{1}{6},$	$\sum_{k,l=1}^{s} b_k a_{kl} c_k = \frac{1}{6}$

2.2 Convex combinations of first- and high-order IMEX schemes

It is well known that approximating discontinuous solutions with high-order non-TVD methods can lead to spurious artefacts near jump positions. This behaviour is illustrated in Figure 1, where we display the approximation of an advected rectangular bump profile with the first-order scheme

$$w_j^{n+1} = w_j^n - \lambda \Delta_j^n - \mu_\varepsilon \Delta_j^{n+1}, \tag{2.7}$$

as well as the well-known second-order ARS(2,2,2) and third-order ARS(2,3,3) IMEX schemes from [2]. For more details on the numerical experiment, such as initial and boundary conditions, see

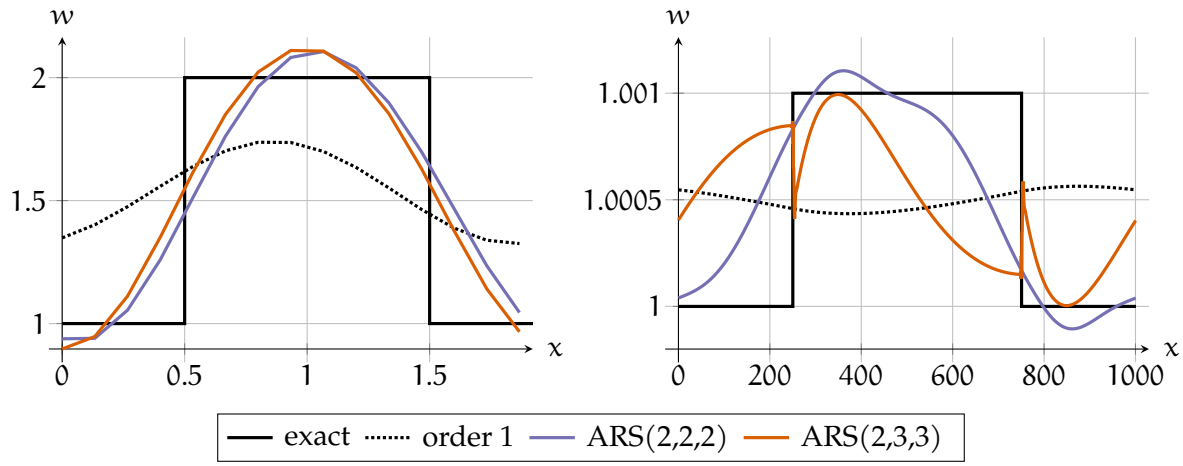

Figure 1: Approximation of a discontinuous solution using the first-order, second-order ARS(2,2,2) and third-order ARS(2,3,3) schemes. Left panel: $\varepsilon = 1$ and $N = 15$; right panel: $\varepsilon = 10^{-3}$ and $N = 2000$. In both cases, the higher-order approximations are oscillatory and the first-order one is diffusive. For more detail on the numerical experiment, see section 5.

section 5. We clearly observe in Figure 1 that the higher-order non-TVD schemes present spurious oscillations in the numerical solution.

Therefore, in order to avoid oscillations as in Figure 1, we need L^∞ stable or TVD schemes. A scheme is said to be L^∞ stable if

$$\|w^{n+1}\|_\infty = \max_{j\in\{1,\dots,N\}} |w_j^{n+1}| \leqslant \|w^n\|_\infty, \tag{2.8}$$

and TVD if

$$\mathrm{TV}(w^{n+1}) = \sum_{j=1}^{N} \left| w_{j+1}^{n+1} - w_j^{n+1} \right| \leqslant \mathrm{TV}(w^n). \tag{2.9}$$

Unfortunately, it can be proven for IMEX RK schemes, following a result of Gottlieb et al. [15], that there cannot exist L^∞ stable IMEX RK schemes of order $p \geqslant 2$ whose CFL restriction only stems from the explicitly treated part. For completeness, we have added the proof in Appendix A. It immediately follows that it does not make sense to look for higher order TVD IMEX integrators. Turning again to Figure 1, we see that the first-order scheme is very diffusive and not practical in this context. Therefore, our main focus here is to construct a first-order IMEX integration scheme fulfilling the L^∞ stability (2.8) and TVD property (2.9), that has a reduced numerical diffusion compared to the first order scheme (2.7).

To achieve this, we propose a convex combination of (2.3) and the first-order scheme (2.7), following [9]. The new update with the parameter $\theta \in [0,1]$ is then given by

$$w_j^{n+1} = w_j^n - \theta \left(\lambda \sum_{k=1}^{s} \tilde{b}_k \Delta_j^{(k)} + \mu_\varepsilon \sum_{k=1}^{s} b_k \Delta_j^{(k)} \right) - (1-\theta) \left(\lambda \Delta_j^n + \mu_\varepsilon \Delta_j^{n+1} \right). \tag{2.10}$$

We emphasise that a TVD scheme resulting from the above given convex combination (2.10) is only first-order accurate due to Proposition 10, but will have a higher resolution, governed by the value of θ, than the usual first-order scheme (2.7).

3 L^∞ stable and TVD scheme based on second-order tableaux

The goal of this section is to provide a theoretical framework to construct L^∞ stable and TVD discretisations from a second-order Butcher tableau based on the general form (2.5). First, we discuss the stability properties of the convex combination scheme (2.10), with respect to the convex combination parameter θ. Then, we propose a strategy to increase the resolution in space of the resulting scheme.

3.1 TVD time integration

We apply the first- and second-order conditions from Table 1 and (2.6) on the Butcher tableaux given in (2.5) with $s = 3$ stages. To reduce the number of effective computational steps s_{eff}, we assume in addition that the weights $\tilde{b}$ and b respectively coincide with the last rows of $\tilde{A}$ and A. This and the CK type structure lead to $s_{\text{eff}} = 2$ and to the following Butcher tableaux, where $\beta \neq \{0, 1\}$:

$$\text{explicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \beta & \beta & 0 & 0 \\ 1 & 1-\frac{1}{2\beta} & \frac{1}{2\beta} & 0 \\ \hline & 1-\frac{1}{2\beta} & \frac{1}{2\beta} & 0 \end{array}, \qquad \text{implicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \beta & 0 & \beta & 0 \\ 1 & 0 & \frac{1}{2(1-\beta)} & 1-\frac{1}{2(1-\beta)} \\ \hline & 0 & \frac{1}{2(1-\beta)} & 1-\frac{1}{2(1-\beta)} \end{array}. \tag{3.1}$$

Due to the particular structure, we can immediately set $w^{(3)} = w^{n+1}$ and $w^{(1)} = w^n$, which does not contribute to the number of effective computational steps. Using the stages given in (2.4), and the convex update (2.10), the scheme is given by

$$w_j^{(2)} + \mu_\varepsilon a_{22} \Delta_j^{(2)} = w_j^n - \lambda \tilde{a}_{21} \Delta_j^n, \tag{3.2a}$$

$$w_j^{n+1} + \mu_\varepsilon \left((1-\theta) + \theta a_{33}\right) \Delta_j^{n+1} = w_j^n - \lambda \left((1-\theta) + \theta \tilde{a}_{31}\right) \Delta_j^n - \theta \left(\lambda \tilde{a}_{32} + \mu_\varepsilon a_{32}\right) \Delta_j^{(2)}. \tag{3.2b}$$

Note that $\tilde{a}_{21} = a_{22}$. We rewrite the ε dependent term in (3.2a) as

$$-\mu_\varepsilon \Delta_j^{(2)} = \frac{1}{a_{22}} \left(w_j^{(2)} - w_j^n\right) + \lambda \Delta_j^n.$$

Using (3.1) in (3.2b) and $\Delta_j^{(\cdot)} = w_j^{(\cdot)} - w_{j-1}^{(\cdot)}$, we find

$$w_j^{(2)} + \mu_\varepsilon a_{22} \Delta_j^{(2)} = (1 - \lambda a_{22}) w_j^n + \lambda a_{22} w_{j-1}^n, \tag{3.3a}$$

$$\begin{aligned} w_j^{n+1} + \mu_\varepsilon \left(1 + \theta(a_{33} - 1)\right) \Delta_j^{n+1} = &\left(1 - \lambda(1 + \theta(\tilde{a}_{31} - a_{32} - 1)) - \frac{\theta a_{32}}{a_{22}}\right) w_j^n \\ &+ \lambda \left(1 + \theta(\tilde{a}_{31} - a_{32} - 1)\right) w_{j-1}^n \\ &+ \theta \left(\frac{a_{32}}{a_{22}} - \lambda \tilde{a}_{32}\right) w_j^{(2)} + \theta \lambda \tilde{a}_{32} w_{j-1}^{(2)}. \end{aligned} \tag{3.3b}$$

In total, we have three free parameters to be fixed, namely $\beta \neq \{0, 1\}$, $\lambda \geqslant 0$ and $\theta \in [0, 1]$. By setting $\beta \in (0, 1)$ with $\lambda < 1$ and $\theta \leqslant 2\beta(1-\beta)$, we find that all coefficients in front of $w_j^{(k)}, w_{j-1}^{(k)}$ on the right-hand sides of (3.3a) and (3.3b) are greater than or equal to zero. In (3.3a), we find:

$$w_j^n : 1 - \lambda a_{22} = 1 - \lambda\beta \geqslant 0, \qquad w_{j-1}^n : \ \lambda a_{22} = \lambda\beta > 0, \tag{3.4}$$

and, in (3.3b), we find

$$\begin{aligned}
w_j^n: &\quad (1-\lambda(1+\theta(\tilde{a}_{31}-a_{32}-1)))-\frac{\theta a_{32}}{a_{22}}=\left(1-\frac{\theta}{2\beta(1-\beta)}\right)-\lambda\left(1-\frac{\theta}{2\beta(1-\beta)}\right)\geqslant 0\\
w_{j-1}^n: &\quad \lambda\left(1+\theta(\tilde{a}_{31}-a_{32}-1)\right)=\lambda\left(1-\frac{\theta}{2\beta(1-\beta)}\right)\geqslant 0\\
w_j^{(2)}: &\quad \theta\left(\frac{a_{32}}{a_{22}}-\lambda\tilde{a}_{32}\right)=\theta\left(\frac{1}{2\beta(1-\beta)}-\lambda\frac{1}{2\beta}\right)>0\\
w_{j-1}^{(2)}: &\quad \theta\lambda\tilde{a}_{32}=\theta\lambda\frac{1}{2\beta}>0.
\end{aligned}\tag{3.5}$$

In addition, since μ_ε is non-negative, we have positive coefficients in front of $\Delta_j^{(2)}$ and Δ_j^{n+1} on the left hand side of equations (3.3a) and (3.3b). With the same notation as above we find

$$\Delta_j^{(2)}: \mu_\varepsilon a_{22}=\beta\mu_\varepsilon>0, \qquad \Delta_j^{n+1}: \mu_\varepsilon\left(1+\theta(a_{33}-1)\right)=\mu_\varepsilon\left(1-\frac{\theta}{2(1-\beta)}\right)>0. \tag{3.6}$$

The inequalities (3.4), (3.5) and (3.6) are the key element to show the L^∞ stability and TVD properties of scheme (3.2), because this ensures the proof only by using the triangle inequality $\|ax+by\|\leqslant a\|x\|+b\|y\|$ and reverse triangle inequality $a\|x\|-b\|y\|\leqslant\|ax-by\|$ for $x,y\in\mathbb{R}$ and $a,b\in\mathbb{R}$ with $a,b\geqslant 0$. We start with the L^∞ stability and show first that $\|w^{(2)}\|_\infty\leqslant\|w^n\|_\infty$. For periodic boundary conditions, we find with (3.4) and (3.6) that

$$\begin{aligned}
\|w^n\|_\infty &= (1-\lambda a_{22})\,\|w^n\|_\infty+\lambda a_{22}\|w^n\|_\infty\\
&= (1-\lambda a_{22})\ \max_j|w_j^n|+\lambda a_{22}\ \max_j|w_{j-1}^n|\\
&\geqslant \max_j\left|(1-\lambda a_{22})\,w_j^n+\lambda a_{22}w_{j-1}^n\right|\\
&= \max_j\left|w_j^n-\lambda a_{22}(w_j^n-w_{j-1}^n)\right|\\
&= \max_j\left|(1+\mu_\varepsilon a_{22})w_j^{(2)}-\mu_\varepsilon a_{22}w_{j-1}^{(2)}\right|\\
&\geqslant (1+\mu_\varepsilon a_{22})\|w^{(2)}\|_\infty-\mu_\varepsilon a_{22}\|w^{(2)}\|_\infty\\
&= \|w^{(2)}\|_\infty.
\end{aligned}$$

Using (3.5) and (3.6), as well as the above estimate $\|w^{(2)}\|_\infty\leqslant\|w^n\|_\infty$, we can also prove analogously

$$\|w^{n+1}\|_\infty\leqslant\left(1-\frac{\theta a_{32}}{a_{22}}\right)\|w^n\|_\infty+\frac{\theta a_{32}}{a_{22}}\|w^{(2)}\|_\infty\leqslant\|w^n\|_\infty.$$

Thus, we have proven the L^∞ stability. We summarize the result in the following lemma.

Lemma 1. *For periodic boundary conditions under the CFL condition $\lambda<1$, the scheme consisting of the Butcher tableaux (3.1) with the convex update (2.10) and the stages (2.4) with the parameters $\beta\in(0,1)$ and $\theta\leqslant 2\beta(1-\beta)$ is L^∞ stable.*

In addition, if the optimal value for θ is taken, that is if $\theta=\theta_{opt}=2\beta(1-\beta)$, then the CFL condition relaxes to $\lambda\leqslant\min\left(\frac{1}{\beta},\frac{1}{1-\beta}\right)$.

Using the same arguments as for the proof of the L^∞ stability, we now show the TVD property. Assuming periodic boundary conditions, we write

$$\mathrm{TV}(w^n)=(1-\lambda a_{22})\sum_{j=1}^N\left|w_{j+1}^n-w_j^n\right|+\lambda a_{22}\sum_{j=1}^N\left|w_j^n-w_{j-1}^n\right|$$

$$
\begin{aligned}
&= \sum_{j=1}^{N} \left(\left| (1-\lambda a_{22})\, w_{j+1}^{n} - (1-\lambda a_{22})\, w_{j}^{n} \right| + \left| \lambda a_{22} w_{j}^{n} - \lambda a_{22} w_{j-1}^{n} \right| \right) \\
&\geqslant \sum_{j=1}^{N} \left| \left((1-\lambda a_{22})\, w_{j+1}^{n} - \lambda a_{22} w_{j}^{n} \right) - \left((1-\lambda a_{22})\, w_{j}^{n} - \lambda a_{22} w_{j-1}^{n} \right) \right| \\
&= \sum_{j=1}^{N} \left| \left((1+\mu_\varepsilon a_{22})\, w_{j+1}^{(2)} - \mu_\varepsilon a_{22} w_{j}^{(2)} \right) - \left((1+\mu_\varepsilon a_{22})\, w_{j}^{(2)} - \mu_\varepsilon a_{22} w_{j-1}^{(2)} \right) \right| \\
&\geqslant \sum_{j=1}^{N} \left(\left| (1+\mu_\varepsilon a_{22}) \left(w_{j+1}^{(2)} - w_{j}^{(2)} \right) \right| - \left| \mu_\varepsilon a_{22} \left(w_{j}^{(2)} - w_{j-1}^{(2)} \right) \right| \right) \\
&= (1+\mu_\varepsilon a_{22}) \sum_{j=1}^{N} \left| w_{j+1}^{(2)} - w_{j}^{(2)} \right| - \mu_\varepsilon a_{22} \sum_{j=1}^{N} \left| w_{j}^{(2)} - w_{j-1}^{(2)} \right| \\
&= \mathrm{TV}(w^{(2)}).
\end{aligned}
$$

Using the above estimate, we now show the final TVD property. Since the proof is straightforward, we write the last step below:

$$
\mathrm{TV}(w^{n+1}) \leqslant \left(1 - \frac{\theta a_{32}}{a_{22}} \right) \mathrm{TV}(w^{n}) + \frac{\theta a_{32}}{a_{22}} \mathrm{TV}(w^{(2)}) \leqslant \mathrm{TV}(w^{n}).
$$

This result is summarized as follows

Lemma 2. *For $\beta \in (0,1)$ and periodic boundary conditions, under the CFL condition $\lambda < 1$ for $\theta \leqslant 2\beta(1-\beta)$, and under the relaxed CFL condition $\lambda \leqslant \min\left(\frac{1}{\beta}, \frac{1}{1-\beta}\right)$ for $\theta = \theta^{opt} = 2\beta(1-\beta)$, the scheme consisting of the Butcher tableaux (3.1) with the convex update (2.10) and the stages (2.4) is TVD.*

Note that for the proof of the Lemmata 1 and 2 we only used the positivity restrictions (3.4), (3.5) and (3.6), as well as the choice of the boundary conditions. This means that the TVD property will always hold under the exact same constraints as the L^∞ stability. Furthermore, the proof holds also for Neumann boundary conditions.

3.2 TVD reconstruction in space

To increase the resolution of the spatial derivatives, we seek a second-order reconstruction of the point values w_j such that the resulting scheme is still L^∞ stable and TVD. We start with the explicit space derivatives.

Explicit space reconstruction. To obtain a second-order accurate approximation of the explicit spatial derivatives, we linearly reconstruct the values $w_j^{(k)}$ using the neighbouring point values, see for instance [24]. The reconstructed values $w_{j,-}^{(k)}$ and $w_{j,+}^{(k)}$ are then defined by

$$
w_{j,-}^{(k)} = w_{j}^{(k)} - \frac{\Delta x}{2} L\left(\sigma_{j+1/2}^{(k)}, \sigma_{j-1/2}^{(k)} \right), \quad w_{j,+}^{(k)} = w_{j}^{(k)} + \frac{\Delta x}{2} L\left(\sigma_{j-1/2}^{(k)}, \sigma_{j+1/2}^{(k)} \right), \tag{3.7}
$$

where $\sigma_{j+1/2}^{(k)}$ denotes the slope between the values of $w_j^{(k)}$ and $w_{j+1}^{(k)}$ given by

$$
\sigma_{j+1/2}^{(k)} = \frac{w_{j+1}^{(k)} - w_{j}^{(k)}}{\Delta x}.
$$

The function $L(\sigma_L, \sigma_R)$ is a slope limiter which should ensure that the reconstructed values still satisfy the maximum principle. For a three-point stencil the following estimate has to hold

$$\min(|w_{j-1}^{(k)}|, |w_j^{(k)}|, |w_{j+1}^{(k)}|) \leqslant |w_{j,\pm}^{(k)}| \leqslant \max(|w_{j-1}^{(k)}|, |w_j^{(k)}|, |w_{j+1}^{(k)}|). \tag{3.8}$$

A popular example of a second-order TVD slope limiter is the minmod limiter, defined for any two slopes σ_L and σ_R by

$$\text{minmod}(\sigma_L, \sigma_R) = \begin{cases} \min(\sigma_R, \sigma_L) & \text{if } \sigma_R > 0 \text{ and } \sigma_L > 0, \\ \max(\sigma_R, \sigma_L) & \text{if } \sigma_R < 0 \text{ and } \sigma_L < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

Using the reconstruction (3.7) and the notation $\Delta_{j,+}^{(k)} = w_{j,+}^{(k)} - w_{j-1,+}^{(k)}$, we write the stages and the update given in (3.2) as

$$\begin{aligned} w_j^{(2)} + \mu_\varepsilon a_{22} \Delta_j^{(2)} &= w_j^n - \lambda \tilde{a}_{21} \Delta_{j,+}^n, \\ w_j^{n+1} + \mu_\varepsilon \left((1-\theta) + \theta a_{33} \right) \Delta_j^{n+1} &= w_j^n - \lambda \left((1-\theta) + \theta \tilde{a}_{31} \right) \Delta_{j,+}^n - \theta \left(\lambda \tilde{a}_{32} + \mu_\varepsilon a_{32} \right) \Delta_{j,+}^{(2)}. \end{aligned}$$

Due to the minmod limiting procedure, we immediately have from the estimate (3.8) that

$$\max_j \left| w_{j,+}^n \right| \leqslant \max_j \left| w_j^n \right| \quad \text{and} \quad \max_j \left| w_{j,+}^{(2)} \right| \leqslant \max_j \left| w_j^{(2)} \right|$$

for periodic boundary conditions. Using this estimates and following the analogue steps in the proofs of Lemma 1 and 2 it is easy to see, that under this reconstruction, the L^∞ stability and TVD property still hold.

Implicit space reconstruction. In the spirit of the reconstruction used to approximate the explicit derivatives, we could also increase the space accuracy of the implicit derivatives using TVD slope limiters. Note that the slopes are determined in general by a non-linear function, for example the minmod limiter (3.9). This would mean having to implicitly compute the reconstructed values (3.7). Such computations, if at all doable, would include an iterative process or a prediction correction method and therefore be extremely costly. We consider this increase in computational cost as too much in the sight of the actual gain in resolution.

Treating the implicit spacial derivative with a BDF to obtain a high-order approximation is not an option here as it leads to oscillatory solutions, see Appendix B for a proof of this claim. Therefore, we keep the first-order upwind approximation of the implicit spatial derivatives. This is a loss of resolution in space we are willing to take to obtain a TVD scheme.

We summarize the results of this section in the following result:

Theorem 3. *For $\beta \in (0, 1)$ and periodic boundary conditions, the scheme consisting of the Butcher tableaux (3.1) with the convex update (2.10) and the stages (2.4), combined with the reconstruction procedure given by (3.7) and (3.9), is L^∞ stable and TVD under the CFL condition $\lambda < 1$ for $\theta \leqslant 2\beta(1-\beta)$, and the relaxed CFL condition $\lambda \leqslant \min(\frac{1}{\beta}, \frac{1}{1-\beta})$ for $\theta = \theta^{opt} = 2\beta(1-\beta)$.*

4 Extension to higher order tableaux

We start the construction of schemes using higher order tableaux by investigating the natural extension of the TVD scheme using third-order tableaux instead of second-order ones. Unfortunately,

as we prove now, a straightforward extension using the same assumptions as in (3.1) does not lead to a TVD scheme. The Butcher tableaux with $s_{\text{eff}} = 3$ effective computational steps are given by

$$\text{explicit:}\quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & \tilde{a}_{21} & 0 & 0 \\ c_3 & \tilde{a}_{31} & \tilde{a}_{32} & 0 \\ \hline & 0 & b_2 & b_3 \end{array}, \qquad \text{implicit:}\quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ c_2 & 0 & a_{22} & 0 \\ c_3 & 0 & a_{32} & a_{33} \\ \hline & 0 & b_2 & b_3 \end{array}. \tag{4.1}$$

We have assumed that $\tilde{b} = b$ and $\tilde{c} = c$ for simplicity, see also [31]. This also has the advantage of updating the whole flux.

Applying the third-order conditions given in Table 1 and (2.6) on the scheme given by (4.1) leads to the following tableaux, with $\gamma \notin \{0, \frac{1}{3}\}$:

$$\text{explicit:}\quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{3\gamma-1}{6\gamma} & \frac{3\gamma-1}{6\gamma} & 0 & 0 \\ \frac{\gamma+1}{2} & -\frac{6\gamma^3-3\gamma^2+1}{2(3\gamma-1)} & \frac{\gamma(3\gamma^2+1)}{3\gamma-1} & 0 \\ \hline & 0 & \frac{3\gamma^2}{3\gamma^2+1} & \frac{1}{3\gamma^2+1} \end{array} \qquad \text{implicit:}\quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{3\gamma-1}{6\gamma} & 0 & \frac{3\gamma-1}{6\gamma} & 0 \\ \frac{\gamma+1}{2} & 0 & \gamma & \frac{1-\gamma}{2} \\ \hline & 0 & \frac{3\gamma^2}{3\gamma^2+1} & \frac{1}{3\gamma^2+1} \end{array} \tag{4.2}$$

We now derive conditions on $\gamma \notin \{0, \frac{1}{3}\}$, $\lambda > 0$ and $\theta \in [0, 1]$ such that the scheme given by (4.2) and the convex combination with the first-order scheme (2.7) is L^∞ stable and TVD. From the first stage, we have $w^{(1)} = w^n$. The second stage with $c_2 = \frac{3\gamma-1}{6\gamma}$ is given by

$$w_j^{(2)} + \mu_\varepsilon c_2 \Delta_j^{(2)} = (1 - \lambda c_2) w_j^n - \lambda c_2 w_{j-1}^n.$$

Following the proof of Lemmata 1 and 2, we require, in the fashion of (3.4)-(3.6):

$$\begin{aligned} c_2 > 0 &\iff \frac{3\gamma-1}{6\gamma} > 0 \iff \gamma < 0 \text{ or } \gamma > \frac{1}{3}, \\ 1 - \lambda c_2 \geqslant 0 &\iff \lambda \leqslant \frac{1}{c_2} \iff \lambda \leqslant \frac{6\gamma}{3\gamma-1}. \end{aligned} \tag{4.3}$$

Note that the expressions in (4.3) are well-defined. The third stage, using

$$-\mu_\varepsilon \Delta_j^{(2)} = \frac{1}{c_2}(w_j^{(2)} - w_j^n) + \lambda \Delta_j^n,$$

is given by

$$w_j^{(3)} + \mu_\varepsilon a_{33} \Delta_j^{(3)} = \left(1 - \frac{a_{32}}{c_2} - \lambda \tilde{a}_{31} + \lambda a_{32}\right) w_j^n + \lambda\left(\tilde{a}_{31} - a_{32}\right) w_{j-1}^n + \left(\frac{a_{32}}{c_2} - \lambda \tilde{a}_{32}\right) w_j^{(2)} + \lambda \tilde{a}_{32} w_{j-1}^{(2)}.$$

This leads to the following inequalities

$$\begin{aligned} \tilde{a}_{32} \geqslant 0 &\iff \frac{\gamma(3\gamma^2+1)}{3\gamma-1} \geqslant 0 \iff \gamma \leqslant 0 \text{ or } \gamma > \frac{1}{3}, \\ \frac{a_{32}}{c_2} - \lambda \tilde{a}_{32} \geqslant 0 &\iff \lambda \leqslant \frac{6\gamma^2}{\gamma(3\gamma^2+1)} \text{ and } \gamma > \frac{1}{3}, \\ \tilde{a}_{31} - a_{32} \geqslant 0 &\iff -\frac{(\gamma+1)(6\gamma^2-3\gamma+1)}{2(3\gamma-1)} \geqslant 0 \iff \gamma \geqslant -1 \text{ and } \gamma < \frac{1}{3}. \end{aligned} \tag{4.4}$$

We remark no γ can fulfil the last two inequalities. Therefore, it is not possible to prove the TVD property following the proof of Lemma 1 and Lemma 2.

In the following, we propose a method to cure this defect and still keep the easy way of proving the TVD property.

4.1 Method of convex stages

As we have seen, the attempt to prove the L^∞ stability and TVD property already failed at the second step, while the convex combination with the first-order scheme is only applied on the final update. Therefore, we propose a convex combination of each stage with a first-order update at time $t^n + c_k\Delta t$ for the k-th stage. To have only one time level, we set $\tilde{c} = c$. This framework allows for more free parameters $\theta_k \in [0, 1]$, where $k = 1, \dots, s$ denotes the stage in the IMEX scheme. To have the best precision possible, the goal is to choose as many θ_k as possible equal to one. Analogously to the convex update (2.10), the stages are given by

$$w_j^{(k)} + (1-\theta_k)c_k\mu_\varepsilon\Delta_j^{(k)} = w_j^n - \lambda\left((1-\theta_k)\tilde{c}_k\Delta_j^n + \theta_k\sum_{l=1}^{k-1}\tilde{a}_{kl}\Delta_j^{(l)}\right) - \mu_\varepsilon\theta_k\sum_{l=1}^{k}a_{kl}\Delta_j^{(l)}. \quad (4.5)$$

Note that, for the Butcher tableaux in the manner of (4.1), we immediately set $\theta_1 = 1$ to recover $w^{(1)} = w^n$. This means the convex stages appear earliest for $k = 2$. In the case where the weights $\tilde{b}$ and b respectively coincide with the last row of $\tilde{A}$ and A, the stage $w^{(s)}$ coincides with the final update w^{n+1}. In particular, we then have $\theta = \theta_s$.

In the spirit of the results from the second-order scheme, we seek a general framework on how to obtain TVD schemes with s stages using the IMEX formulation (2.10) – (4.5). Since the proof follows analogue steps as in Lemmata 1 and 2, we do not repeat the calculations and we directly give the final result.

Theorem 4. *Let $\tilde{A}, A \in \mathbb{R}^{s\times s}$, $\tilde{b}, b, \tilde{c}, c \in \mathbb{R}^s$ define two Butcher tableaux (2.5) fulfilling (2.6) and the p-th order compatibility conditions. Let $\tilde{b}$ and b coincide with the last rows of $\tilde{A}$ and A respectively. For $k = 1, \dots, s$ and $l = 1, \dots, k-1$, we define*

$$\mathcal{A}_k = \theta_k a_{kk} + (1-\theta_k)c_k, \quad \tilde{\mathcal{A}}_k = \theta_k a_{k1} + (1-\theta_k)\tilde{c}_k, \quad \mathcal{B}_{kl} = \frac{\theta_k a_{kl}}{\mathcal{A}_l}, \quad \tilde{\mathcal{B}}_{kl} = \theta_k\tilde{a}_{kl}.$$

In addition, we recursively define the following expressions:

$$\mathcal{C}_k = \tilde{\mathcal{A}}_k - \sum_{l=2}^{k-1}\mathcal{B}_{kl}\mathcal{C}_l, \qquad \mathcal{C}_{kl} = \tilde{\mathcal{B}}_{kl} - \sum_{r=l+1}^{k-1}\mathcal{B}_{kr}\mathcal{C}_{rl},$$

$$\mathcal{D}_k = 1 - \lambda\tilde{\mathcal{A}}_k - \sum_{l=2}^{k-1}\mathcal{B}_{kl}\mathcal{D}_l, \qquad \mathcal{D}_{kl} = \mathcal{B}_{kl} - \lambda\tilde{\mathcal{B}}_{kl} - \sum_{r=l+1}^{k-1}\mathcal{B}_{kr}\mathcal{D}_{rl}.$$

Then, with $\theta_1 = 1$ and under the following restrictions for $k = 2, \dots, s$ and $l = 1, \dots, k-1$,

$$\mathcal{A}_k > 0, \quad \mathcal{C}_k \geqslant 0, \quad \mathcal{D}_k \geqslant 0, \quad \mathcal{C}_{kl} \geqslant 0, \quad \mathcal{D}_{kl} \geqslant 0.$$

the scheme consisting of the stages (4.5) and the update (2.10), combined with a TVD limiter, is L^∞ stable and TVD under a CFL condition determined by $\lambda \geqslant 0$ where λ does not depend on ε.

We wish to remark that the obtained p-th order tableaux do not necessarily lead to stable schemes by themselves if they are not combined with the convex strategy. This is not a drawback since our goal is the L^∞ stability. For studies on A- or L-stability, we refer to [30].

The result from Theorem 4 can be extended to the case where the weights $\tilde{b}$ and b do not coincide with the respective last rows of $\tilde{A}$ and A. To be able to use the notation from Theorem 4, we view the update (2.10) as an additional explicit $(s+1)$-th stage of a scheme induced by Butcher tableaux (2.5) with $(s+1)\times(s+1)$ matrices with the diagonal entry $a_{s+1,s+1} = 0$, where the weights $\tilde{b}$ and b respectively coincide with the last rows of the new $\tilde{A}$ and A. Then we define the convex parameter of the last stage as $\theta_{s+1} = \theta$. Theorem 4 is then applied to yield the L^∞ stability and the TVD property.

Remark 5. *We can prove the same kind of theorem if the first column of A allows for non-zero entries. The TVD conditions obtained when assuming that structure are given in Appendix C.*

4.2 L^∞ stable and TVD scheme based on third-order tableaux

We now demonstrate that, with this method, a TVD scheme can be obtained based on the previous Butcher tableaux (4.2). Let us introduce one additional parameter $\theta_3 \neq 1$, while keeping $\theta_1 = \theta_2 = 1$. This means that we have the same stages for $w^{(1)}$ and $w^{(2)}$ as before. We recall that we obtained from the second stage $\gamma < 0$ or $\gamma > \frac{1}{3}$ and $\lambda \leqslant \frac{6\gamma}{3\gamma-1}$. Using the definition of the third stage given in (4.5), we now have with $w^{(1)} = w^n$:

$$\begin{aligned} w_j^{(3)} + \mu_\varepsilon \left((1-\theta_3)c_3 + \theta_3 a_{33}\right)\Delta_j^{(3)} &= w_j^n - \lambda\left((1-\theta_3)c_3\Delta^n + \theta_3 c_3\left(\tilde{a}_{31}\Delta_j^n + \tilde{a}_{32}\Delta^{(2)}\right)\right) \\ &\quad + \theta_3 c_3 a_{32}\left(\frac{1}{c_2}(w_j^{(2)} - w_j^n) + \lambda\Delta_j^n\right) \\ \Longleftrightarrow & \\ w_j^{(3)} + \mu_\varepsilon\left(c_3 + \theta_3(a_{33}-c_3)\right)\Delta_j^{(3)} &= \left(1 - \lambda(1-\theta_3)c_3 - \theta_3\lambda(\tilde{a}_{31}-a_{32}) - \frac{\theta_3 a_{32}}{c_2}\right) w_j^n \\ &\quad + \left(\lambda(1-\theta_3)c_3 + \theta_3\lambda(\tilde{a}_{31}-a_{32})\right) w_{j-1}^n \\ &\quad + \theta_3\left(\frac{a_{32}}{c_2} - \lambda\tilde{a}_{32}\right) w_j^{(2)} + \theta_3\,\lambda\tilde{a}_{32} w_{j-1}^{(2)}. \end{aligned}$$

As in the previous case, we obtain

$$\begin{aligned} \tilde{a}_{32} \geqslant 0 &\iff \frac{\gamma(3\gamma^2+1)}{3\gamma-1} \geqslant 0 \iff \gamma \leqslant 0 \text{ or } \gamma > \frac{1}{3}, \\ \frac{a_{32}}{c_2} - \lambda\tilde{a}_{32} \geqslant 0 &\iff \lambda \leqslant \frac{6\gamma^2}{\gamma(3\gamma^2+1)} \text{ and } \gamma > \frac{1}{3}. \end{aligned}$$

For the requirement (4.4) that caused problems earlier, instead of $\tilde{a_{31}} - a_{32} \geqslant 0$, we get

$$(1-\theta_3)c_3 + \theta_3(\tilde{a_{31}} - a_{32}) \geqslant 0 \iff \theta_3\frac{3\gamma^2(\gamma+1)}{3\gamma-1} \leqslant \frac{\gamma+1}{2} \iff \theta_3 \leqslant \frac{3\gamma-1}{6\gamma^2}, \tag{4.6}$$

thus leading to a restriction on θ_3 instead of on γ. The next restriction gives another estimate on θ_3, as follows:

$$c_3 + \theta_3(a_{33}-c_3) \geqslant 0 \iff \gamma\theta_3 \leqslant \frac{\gamma+1}{2} \iff \theta_3 \leqslant \frac{\gamma+1}{2\gamma}.$$

It is easy to see that this condition on θ_3 is less restrictive than the one obtained from (4.6) for all $\gamma > \frac{1}{3}$. For a given γ, the largest value we can take for θ_3 is therefore given by

$$\theta_3^{\text{opt}} = \frac{3\gamma-1}{6\gamma^2},$$

and θ_3 must satisfy $\theta_3 \leqslant \theta_3^{\text{opt}}$. The last restriction for the third stage is given by

$$1 - \lambda(1-\theta_3)c_3 - \theta_3\lambda(\tilde{a}_{31}-a_{32}) - \frac{\theta_3 a_{32}}{c_2} \geqslant 0. \tag{4.7}$$

This condition is always fulfilled if we choose $\theta_3 = \theta_3^{\text{opt}}$. In doing so, we have the maximal allowed input from the original stages (2.4). Otherwise, (4.7) leads to another, more restrictive estimate for λ. We repeat this procedure for the last stage. We skip the lengthy but straightforward computations and give the final estimates on the free parameters $\gamma, \lambda, \theta_3$ and θ_4 directly in Corollary 6.

Explicit space reconstruction. To increase the space accuracy of the scheme, we use a TVD third-order space reconstruction satisfying (3.8). This merely amounts to setting the limiter function L in the space reconstruction described in section 3.2. We choose the third-order limiting procedure introduced in [33]. This procedure switches between the oscillatory non-limited third-order reconstruction and a third-order TVD limiter. Switching to the TVD limiter is triggered in the event where a non-physical oscillation represented by a non-smooth extremum is detected. Since the limiter is provably TVD according to [33], we apply Theorem 4 and immediately find the following result

Corollary 6. *The scheme consisting of the Butcher tableaux (4.2), with the stages given in (4.5), the update in (2.10), and combined with the slope limiter from [33], is L^∞ stable and TVD according to Theorem 4 with the following choice of parameters*

$$\gamma \geqslant \frac{\sqrt{3}}{3}, \quad \theta_1 = 1, \quad \theta_2 = 1, \quad \theta_3 = \frac{3\gamma - 1}{6\gamma^2}, \quad \theta_4 < \frac{(3\gamma - 1)(3\gamma^2 + 1)}{18\gamma^3},$$

and under the CFL condition

$$\lambda \leqslant \frac{18\gamma^3\theta_4 - (3\gamma - 1)(3\gamma^2 + 1)}{(3\gamma - 1)((6\gamma^2 + 1)\theta_4 - (3\gamma^2 + 1))}.$$

An analysis of the influence of the choice of the parameters $\gamma, \theta_3, \theta_4$ and λ will be conducted in section 5.2. Especially the balance between CPU time, i.e. the value of λ, and precision, expressed by the values of θ_3 and θ_4, will be discussed. We emphasise once again that the bound on the time step, expressed by λ, does not depend on ε, which represents the fast scale in equation (2.1).

5 Numerical results

In this last section, we illustrate the capabilities of the schemes we have developed in Sections 3 and 4. To help referring to these methods, we introduce the following abbreviations.

- The IMEXp scheme denotes the scheme with an p-th order time discretisation and an p-th order space discretisation. Following this notation, the IMEX1 scheme is given by (2.7), the IMEX2 scheme corresponds to the Butcher tableaux (3.1), and the IMEX3 scheme corresponds to the Butcher tableaux (4.2). The second-order unlimited space discretisation (3.7) with $L(\sigma_L, \sigma_R) = \frac{1}{2}(\sigma_L + \sigma_R)$ is applied to the explicit part of the IMEX2 scheme, while the second-order BDF (B.1) is applied to its implicit part. The third-order unlimited space discretisation from [33] is applied to the explicit part of the IMEX3 scheme, while the third-order BDF (B.2) is applied to its implicit part.
- The TVDp scheme is the TVD scheme constructed from the IMEXp tableau. The TVD2 scheme is obtained following Theorem 3, and the TVD3 scheme is given in Corollary 6.

For the remainder of this section, we consider several numerical experiments, with some common characteristics. In each experiment, we prescribe periodic boundary conditions, and we take $c_m = 1$ and $c_a = 1$. The value of the fast transport velocity therefore is $1/\varepsilon$. The values of ε will vary throughout the experiments to highlight how the results depend on ε. The space-time domain is taken such that the solution revolves exactly once with the periodic boundary conditions, i.e. we take the final time $t_{end} = 1$ and space domain $(0, c_m + \frac{c_a}{\varepsilon})$.

We introduce two exact solutions to the initial value problem (2.1), which will help us demonstrate the properties of the schemes. First, we give a smooth solution $w^s(t, x)$ defined by

$$w^s(t, x) = 1 + \frac{\varepsilon}{2}\left(1 + \sin\left[2\pi\varepsilon\left(x - \left(c_m + \frac{c_a}{\varepsilon}\right)t\right)\right]\right), \tag{5.1}$$

which represents a sine function of amplitude ε, transported with the velocity $c_m + \frac{c_a}{\varepsilon}$. Second, a discontinuous solution $w^d(t,x)$ is given by

$$w^d(t,x) = \begin{cases} 1+\varepsilon & \text{if } \frac{1}{4} < \left(\frac{\left(x - \left(c_m + \frac{c_a}{\varepsilon}\right) t\right)}{c_m + \frac{c_a}{\varepsilon}} - \left\lfloor \frac{\left(x - \left(c_m + \frac{c_a}{\varepsilon}\right) t\right)}{c_m + \frac{c_a}{\varepsilon}} \right\rfloor \right) < \frac{3}{4} \\ 1 & \text{otherwise,} \end{cases} \tag{5.2}$$

which represents a rectangular bump of amplitude ε, transported with the velocity $c_m + \frac{c_a}{\varepsilon}$, and initially located in the space region $\left(\frac{1}{4}(c_m + \frac{c_a}{\varepsilon}), \frac{3}{4}(c_m + \frac{c_a}{\varepsilon})\right)$. These exact solutions will be taken as initial conditions by setting $t = 0$.

In the remainder of this section, we first introduce a MOOD procedure to increase the precision of the TVDp scheme in section 5.1. Then, in section 5.2, we study the influence of the free parameters in the TVD2 and TVD3 schemes on the precision and computational time. After having fixed the parameters, we compare in section 5.3 the behaviour of these schemes, for a wide range of ε, to that of IMEX schemes from the literature. We study both the accuracy of the schemes on the smooth solution (5.1), and the overshoot/undershoot magnitude on the discontinuous solution (5.2). The section is concluded with an application to the isentropic Euler equations in section 5.4.

5.1 Optimal order detection: MOOD-inspired procedure

The goal of this section is to introduce a MOOD-like procedure to increase the precision of the TVDp scheme without degrading its stability properties. The usual MOOD framework for explicit schemes, see e.g. [6], consists in locally and gradually lowering the order of the scheme when an oscillation is detected. In our IMEX case, the non-local nature of the implicit part prevents us from only recomputing the approximate solution on a few selected cells, and the solution has to be recomputed on the whole mesh. To avoid a prohibitive increase in the computation time, we instead suggest to directly use the TVDp scheme on the whole mesh as soon as an oscillation is detected in some cell. In addition, we state that an oscillation has been detected if the approximate solution does not satisfy the bounds of the initial condition.

This implicit MOOD framework is summarized in the following algorithm, which has also been stated in [9, 28].

Algorithm 7 (MOODp scheme)**.** *Equipped with the stable TVD*p *scheme, the MOOD*p *scheme consists in applying the following procedure at each time step:*

1. *Compute a candidate numerical solution w_c^{n+1} with the IMEX*p *scheme.*
2. *Detect whether an oscillation is present somewhere in the space domain, that is to say detect whether the discrete maximum principle is satisfied by the candidate solution:*

$$\|w_c^{n+1}\|_\infty \leqslant \|w^0\|_\infty. \tag{DMP}$$

(*3a*) *If* (DMP) *holds, then set the numerical solution w^{n+1} equal to the candidate solution w_c^{n+1}.*

(*3b*) *Otherwise, compute the numerical solution w^{n+1} with the* L^∞ *stable TVD*p *scheme.*

Applied at each time step, the procedure described in Algorithm 7 ensures that the numerical solution satisfies the maximum principle, i.e. $\|w^{n+1}\|_\infty \leqslant \|w^0\|_\infty$ for all $n \geqslant 0$.

5.2 Choice of the free parameters

We start these numerical experiments by suggesting optimal values of the free parameters in the schemes from Sections 3 and 4. To that end, we analyse the error produced by the schemes, as well as the CPU time taken, with respect to the free parameters. This analysis will help us give some insights on how to optimally choose these parameters, and on the trade-offs that must be made when making such choices.

Here, we study the effect of the time discretisation on the precision and computational time of our schemes. Therefore, we temporarily restrict ourselves to a first-order discretisation in space, in order to make sure only the effects of the time discretisation are studied. We compare the IMEX1 scheme to the IMEX2, TVD2 and MOOD2 schemes in section 5.2.1, and to the IMEX3, TVD3 and MOOD3 schemes in section 5.2.2. In both cases, we set $\varepsilon = 0.1$ and we take $N = 400$ discretisation points, and the smooth exact solution (5.1) is considered. The conclusions of the forthcoming developments are unchanged if we consider other values of ε. Indeed, taking a different ε would merely translate the curves without changing their relative positioning. This study is concluded with the results of an optimisation procedure leveraging the conditions of Theorem 4 in order to build new Butcher tableaux, and their accompanying values of λ and θ, that yield TVD schemes.

5.2.1 Choice of β in the TVD2 scheme

We consider the TVD2 scheme. According to Lemma 2, we can freely choose $\beta \in (0, 1)$ and get a TVD scheme as long as $\theta = 2\beta(1-\beta)$ and $\lambda = \min(\frac{1}{\beta}, \frac{1}{1-\beta})$. These two quantities are displayed in Figure 2. We observe that $\beta = \frac{1}{2}$ maximizes both θ and λ. In this case, the Butcher tableaux (3.1) degenerate to the Butcher tableaux of the ARS (1,2,2) midpoint scheme, see [2], and we get $\theta = \frac{1}{2}$ and $\lambda = 2$. With these settings, the TVD2 scheme exactly reverts to two steps of the IMEX1 scheme, and we expect a loss of accuracy. Therefore, to base the TVD2 scheme on a truly second-order IMEX2 scheme, we have to take $\beta \neq \frac{1}{2}$.

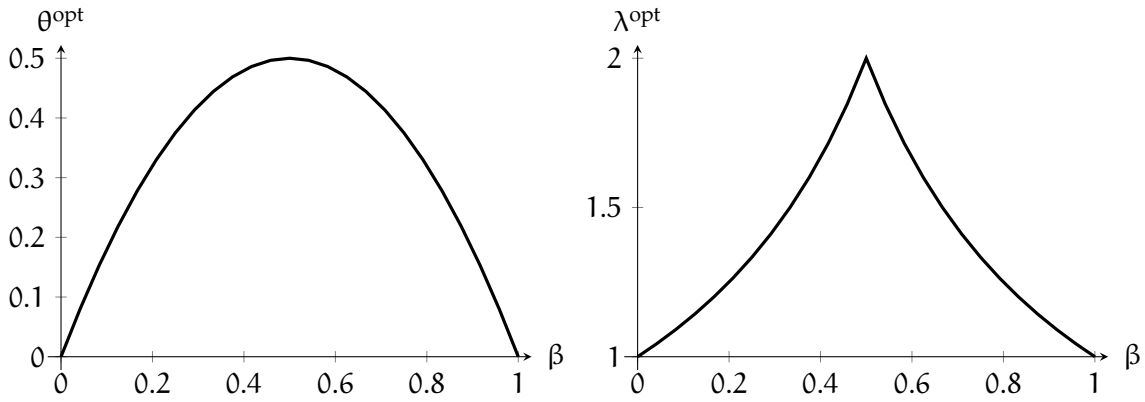

Figure 2: Values of the optimal convex combination parameter θ^{opt} (left panel) and the optimal CFL number λ^{opt} (right panel), with respect to the IMEX parameter β of the TVD2 scheme.

Let us now study the impact of the choice of β on the precision and speed of the numerical scheme. This study was partially performed, for $\beta < \frac{1}{2}$, in [28]. First, we check the CPU time with respect to β. Since the CFL condition of the TVD2 and MOOD2 schemes is influenced by β, we expect these two schemes to take more computational time when β is far from $\frac{1}{2}$. These observations are confirmed by Figure 3.

In the left panel of Figure 4, we display the L^∞-error of the four schemes with respect to β. We observe that the L^∞-error of the IMEX2 scheme explodes around $\beta = 0.52$, even for this smooth solu-

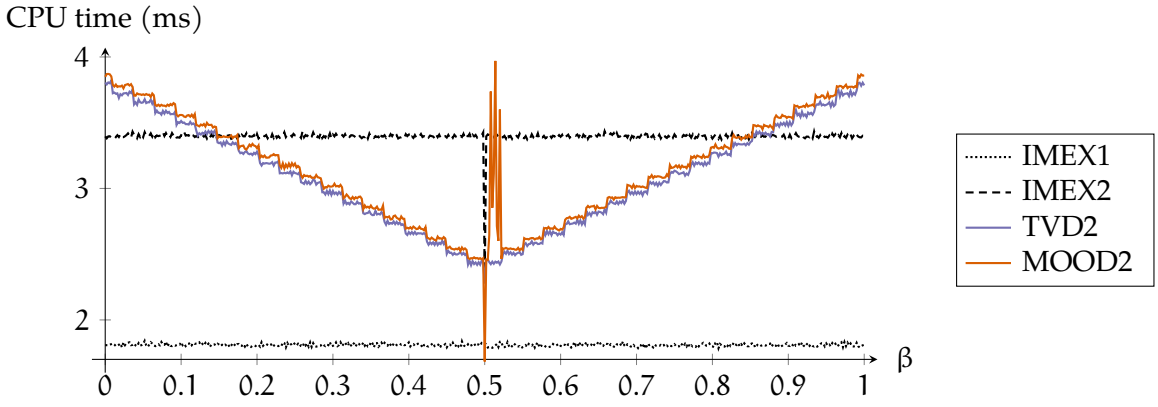

Figure 3: CPU time (in milliseconds) with respect to the IMEX parameter β, using the optimal values θ^{opt} and λ^{opt}, in the context of the test case presented in Section 5.2.1.

tion, which explains the increase in CPU time of the MOOD2 scheme noted in Figure 2. Furthermore, still in the left panel, we observe that the error of both the IMEX2 and the MOOD2 scheme increase sharply when $\beta > \frac{1}{2}$. Therefore, it seems sensible to restrict this study to $\beta < \frac{1}{2}$. In the right panels of Figure 4, we display zooms of the left panel error data for $\beta < \frac{1}{2}$. In the top right panel, we observe that the error of the TVD2 scheme reaches a minimum around $\beta = 0.3$; in the bottom right panel, we observe that the error of the MOOD2 scheme starts increasing around $\beta = 0.3$.

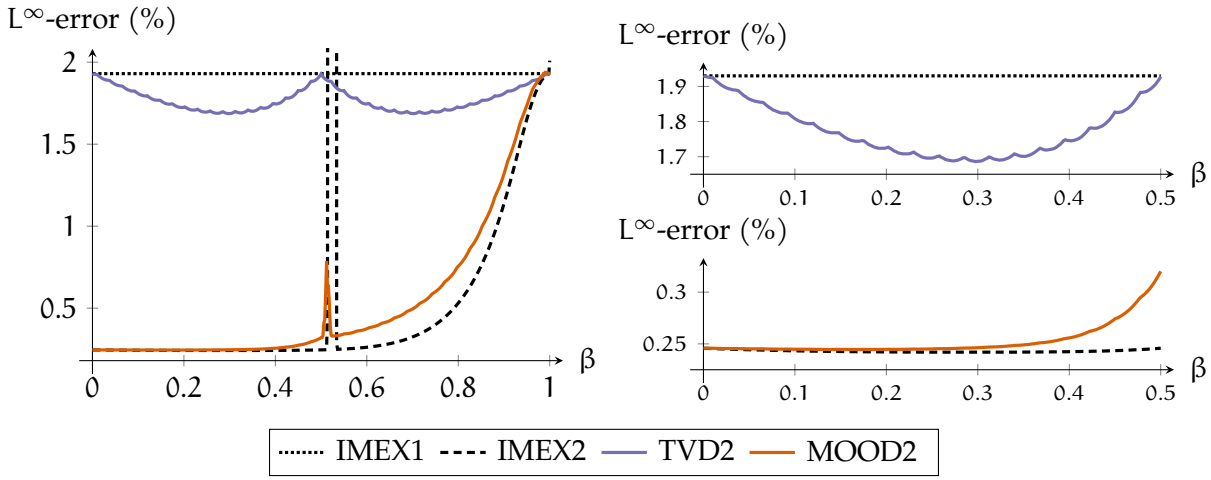

Figure 4: L^∞-error with respect to the IMEX parameter β, using the optimal values θ^{opt} and λ^{opt}, in the context of the test case presented in Section 5.2.1. The right panels contain a zoom on the left panel data, for $\beta \in [0, \frac{1}{2}]$.

Therefore, according to Figures 3 and 4, taking $\beta \simeq 0.29$ seems like a good compromise between error and CPU time taken. We propose $\beta^{opt} = 1 - \frac{\sqrt{2}}{2} \approx 0.293$, leading to the well-known ARS(2,2,2) scheme (see for instance [2, 30]). The Butcher tableaux (3.1) then become

$$\text{explicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1-\frac{\sqrt{2}}{2} & 1-\frac{\sqrt{2}}{2} & 0 & 0 \\ 1 & -\frac{\sqrt{2}}{2} & 1+\frac{\sqrt{2}}{2} & 0 \\ \hline & -\frac{\sqrt{2}}{2} & 1+\frac{\sqrt{2}}{2} & 0 \end{array}, \qquad \text{implicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1-\frac{\sqrt{2}}{2} & 0 & 1-\frac{\sqrt{2}}{2} & 0 \\ 1 & 0 & \frac{\sqrt{2}}{2} & 1-\frac{\sqrt{2}}{2} \\ \hline & 0 & \frac{\sqrt{2}}{2} & 1-\frac{\sqrt{2}}{2} \end{array}.$$

For the remainder of this article, we take

$$\beta = \beta^{\text{opt}} = 1 - \frac{\sqrt{2}}{2}.$$

5.2.2 Choice of γ and θ_4 in the TVD3 scheme

Now, regarding the TVD3 scheme, we have to set the values of θ_3, θ_4 and λ, constrained by Corollary 6. Ideally, we would like θ_3, θ_4 and λ to be as large as possible. By inspection, we note that the maximum value of θ_3 is $\theta_3^{\text{opt}} = \frac{3}{8}$, obtained for $\gamma^{\text{opt}} = \frac{2}{3}$. The Butcher tableaux (4.2) then become

$$\text{explicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 0 & 0 \\ 5/6 & -13/18 & 14/9 & 0 \\ \hline & 0 & 4/7 & 3/7 \end{array}, \qquad \text{implicit:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/4 & 0 & 1/4 & 0 \\ 5/6 & 0 & 2/3 & 1/6 \\ \hline & 0 & 4/7 & 3/7 \end{array}. \tag{5.3}$$

Taking this value of γ in Corollary 6 yields the following bounds:

$$0 < \theta_4 < \frac{7}{16} \quad \text{and} \quad 0 < \lambda < \frac{7 - 16\theta_4}{7 - 11\theta_4}. \tag{5.4}$$

We note that λ is a decreasing function of θ_4, which implies that we are not able to use both a large θ_4 and a large λ. There is a trade-off between the CFL condition λ (i.e. the CPU time) and the value of θ (i.e. the resolution of the scheme).

Let us quantify this balance between precision and CPU time. To address this issue, let us introduce $\alpha \in (0, 1)$, to rewrite (5.4) as follows:

$$\theta_4 = \frac{7}{16}\alpha \quad \text{and} \quad \lambda = \frac{1 - \alpha}{1 - \frac{11}{16}\alpha}. \tag{5.5}$$

In Figure 5, we display the values of θ_4 and λ with respect to α. We indeed note that θ_4 increases and λ decreases when α increases.

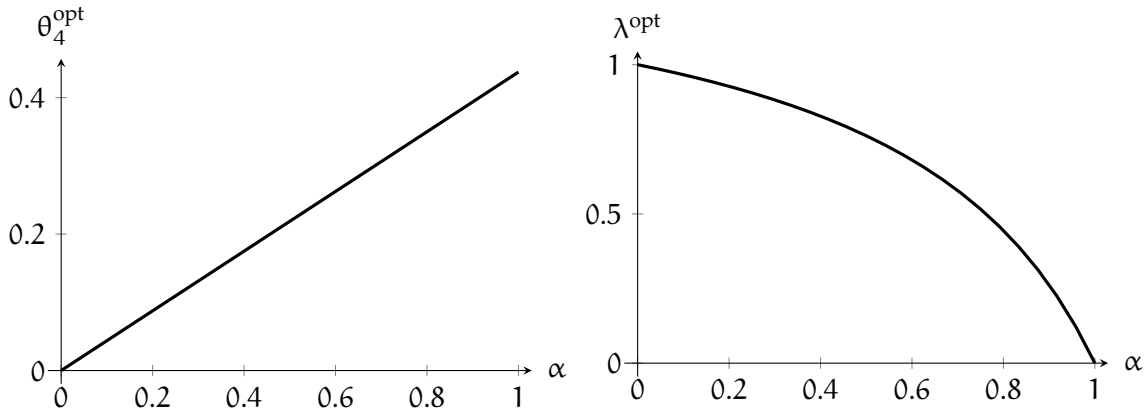

Figure 5: Values of the last convex combination parameter θ_4 (left panel) and the CFL number λ (right panel), with respect to the parameter α, for the TVD3 scheme with $\gamma = \gamma^{\text{opt}} = \frac{2}{3}$.

We now repeat the experiments from Section 5.2.1, this time looking at the influence of α on the TVD3 scheme with $\gamma = \gamma^{\text{opt}} = \frac{2}{3}$. We first display in Figure 6 the CPU time with respect to α for the

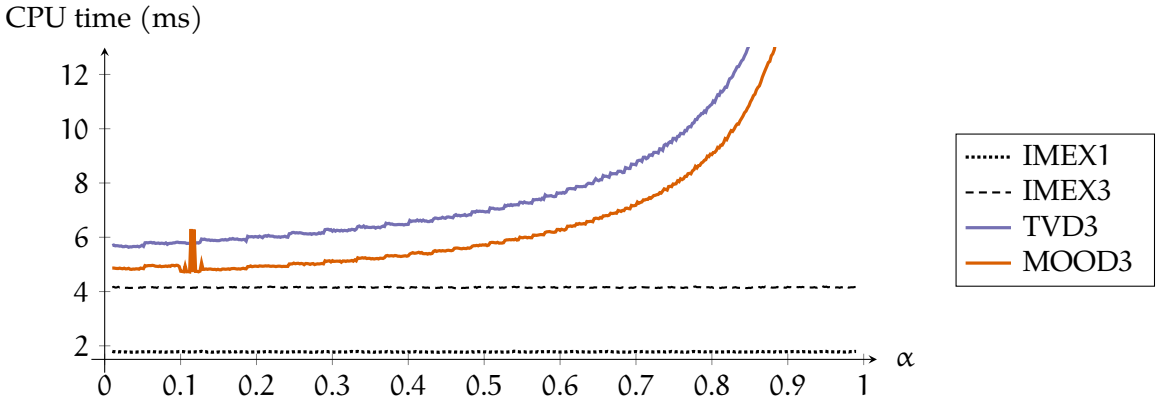

Figure 6: CPU time (in milliseconds) with respect to the parameter α, using $\gamma = \gamma^{opt} = \frac{2}{3}$, in the context of the test case presented in Section 5.2.2.

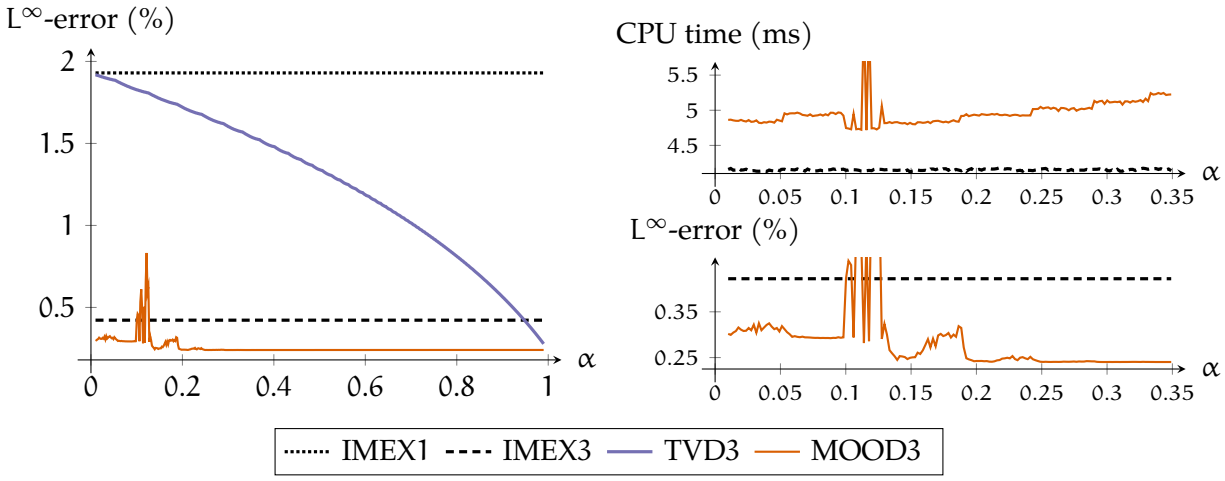

Figure 7: L^∞-error with respect to the parameter α, using $\gamma = \gamma^{opt} = \frac{2}{3}$, $\theta_3 = \frac{3}{8}$ and θ_4, λ given by (5.5). in the context of the test case presented in Section 5.2.2. For $\alpha \in (0, 0.35)$, the top right panel contains a zoom on the CPU time (data from Figure 6 and the bottom right panel contains a zoom on the L^∞-error (data from left panel).

four schemes. As expected, since the CFL condition becomes more restrictive, the CPU time increases with α for the TVD3 and the MOOD3 schemes.

Now, in the left panel of Figure 7, we display the L^∞-error with respect to α for the four schemes under consideration. As expected, we observe that it decreases with α for the TVD3 scheme, since θ_4 increases.

In the right panel of Figure 7, we display a zoom on the CPU time and the L^∞-error produced by the IMEX3 and MOOD3 schemes, with respect to $0 < \alpha < 0.35$. We observe that the error stabilizes around $\alpha = 0.3$, and that the CPU time increases monotonically with α. Therefore, taking $\alpha = \frac{1}{3}$ seems to be a good compromise between precision and computational time. In the remainder of this article, we take

$$\gamma = \gamma^{opt} = \frac{2}{3} \quad \text{and} \quad \alpha = \alpha^{opt} = \frac{1}{3},$$

which leads to the following values for θ_3, θ_4 and λ:

$$\theta_3^{\text{opt}} = \frac{3}{8} = 0.375, \quad \theta_4^{\text{opt}} = \frac{7}{48} \simeq 0.146, \quad \text{and} \quad \lambda^{\text{opt}} = \frac{32}{37} \simeq 0.865.$$

5.2.3 Numerical optimisation of larger Butcher tableaux

To conclude this Section, we mention two other Butcher tableaux that yield a TVD scheme. To obtain these tableaux, we have used the TVD inequalities from Theorem 4, as well as the order conditions from Table 1, as constraints in an optimisation problem where the objective is to maximize the value of $\lambda + \sum \theta$, and where the unknowns are the Butcher coefficients, the values of θ and λ. We ran this optimization problem with many random initial conditions for the unknowns, and we refined this random initialisation around values yielding a large value of the objective function. In the end, we chose the solution where the value of the objective function was maximal, under the additional constraint that $\lambda \geqslant 0.5$ which is a standard CFL condition arising in fluid dynamical schemes.

Three-step, second-order tableau. In this case, we obtain $\lambda = 2.25$, $\theta_1 = 1$, $\theta_2 = 1$, $\theta_3 = 1$ and $\theta_4 = 2/3$ which corresponds to an effective convex combination only in the last stage. The Butcher tableaux are given in Appendix D and in the remainder of the paper, the scheme and its MOOD version will be referred to as TVD2(3) and MOOD2(3).

Four-step, third-order tableau. In this case, we obtain $\lambda = 0.5471076190680170$, $\theta_1 = 1$, $\theta_2 = 1$,$\theta_3 = 1$, $\theta_4 = 0.5110907014643069$ and $\theta_5 = 0.4997722865197203$. The Butcher tableaux are given in Appendix E. In the remainder of the paper, the scheme and its MOOD version will be referred to as TVD3(4) and MOOD3(4).

5.3 Numerical tests and comparison with L-stable and SSP schemes

Now that the optimal values of the free parameters are established, let us test the obtained schemes on a few numerical experiments. We first show, in section 5.3.1, the flexibility of our large time step schemes compared to L-stable and SSP IMEX schemes from the literature. We then check in section 5.3.2 the order of accuracy using the smooth solution (5.1), and we finally study the behaviour of our schemes on the discontinuous solution (5.2) in section 5.3.3. We expect the IMEXp schemes to behave well on smooth solutions, while their non-L^∞ stable nature should produce oscillations and destroy the numerical approximation of discontinuous solutions.

With the choice of β from section 5.2.1, the IMEX2 scheme turns out to be the well-known ARS(2,2,2) scheme. However, the IMEX3 scheme, given by the tableaux (5.3), is not well-known in the literature. To provide a point of comparison, we introduce the L-stable ARS(2,3,3) scheme, reported in [2], Section 2.4, or [30], Table 5, given by the following tableaux

$$\text{expl.:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \delta & \delta & 0 & 0 \\ 1-\delta & \delta-1 & 2-2\delta & 0 \\ \hline & 0 & 1/2 & 1/2 \end{array}, \qquad \text{impl.:} \quad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \delta & 0 & \delta & 0 \\ 1-\delta & 0 & 1-2\delta & \delta \\ \hline & 0 & 1/2 & 1/2 \end{array}, \qquad \text{where } \delta = \frac{3+\sqrt{3}}{6}.$$

Note that this scheme falls within the framework of section 4. Indeed, the above tableaux are nothing but the tableaux (4.2) with $\gamma = \frac{3-\sqrt{3}}{6}$. This value of γ does not satisfy the requirement of Corollary 6, and therefore we cannot prove the existence of convex combinations that make the ARS(2,3,3) scheme

TVD and L^∞ stable with a CFL restriction independent of ε. The following numerical experiments should therefore highlight that the property of L-stability is not enough to ensure non-oscillatory approximations.

Remark 8. *In the following numerical experiments, some values of the number of points* N *are large when* ε *is small. These large values of* N *have been chosen to ensure that more than* 10 *time iterations are needed to reach* t_{end}*. If fewer time iterations are considered, the time steps are too large to visually notice the differences between the schemes.*

5.3.1 On the flexibility of large time step schemes

In this Section, we are concerned with a comparison of the MOOD3(4) with schemes from the literature. Namely, we consider the aforementioned ARS(2,3,3) scheme from [2], as well as a more recent SSP-IMEX scheme from [7], Section 3.2.3, which we label as CGGS3. To compare the results of these two schemes with our MOOD3(4) scheme, we choose to compute both the CPU time and the L^1 error, for the discontinuous solution (5.2) with $N = 4000$ and $\varepsilon = 10^{-3}$.

By inspection, we remark that for the ARS(2,3,3) and CGGS3 schemes to be L^∞ stable, the ε-dependent CFL restriction $\lambda \leqslant 0.9\varepsilon$ is required, while for the MOOD3(4) scheme it is enough to take $\lambda < \lambda^{\text{opt}} \simeq 0.547$.

The results are displayed in Table 2. We observe that the errors and CPU times of the ARS(2,3,3) and CGGS3 schemes are similar for $\lambda = 0.9\varepsilon$. However, in the case of the MOOD3(4) scheme, we can take a much larger range of λ. Indeed, by taking a larger λ, we can choose to sacrifice accuracy and gain CPU time, as evidenced by the first lines of Table 2. Then, taking a smaller λ, for instance $\lambda = 2\varepsilon$, yields a similar CPU time and error compared to the ARS(2,3,3) and CGGS3 schemes.

Table 2: CPU times and L^1 errors for discontinuous solution (5.2) with $N = 4000$ discretisation points and $\varepsilon = 10^{-3}$, using the MOOD3(4), ARS(2,3,3) and CGGS3 schemes.

	λ	CPU time (s)	L^1 error
MOOD3(4)	$\lambda = \lambda^{\text{opt}} \simeq 0.548$	0.0101	0.217
	$\lambda = 250\varepsilon = 0.25$	0.0222	0.111
	$\lambda = 50\varepsilon = 0.05$	0.0953	0.0591
	$\lambda = 10\varepsilon = 0.01$	0.659	0.0488
	$\lambda = 2\varepsilon = 0.0002$	1.63	0.0253
	$\lambda = 0.9\varepsilon = 0.0009$	3.64	0.0253
CGGS3	$\lambda = 0.9\varepsilon = 0.0009$	1.25	0.0253
ARS(2,3,3)	$\lambda = 0.9\varepsilon = 0.0009$	1.17	0.0253

This series of numerical experiments highlights the flexibility of our approach. Based on the desired application and focus where L^∞ stability is required, thanks to large time step schemes, one may choose either a fine resolution following the fastest scale when necessary at the cost of a larger CPU time, or a smaller CPU time when a coarser resolution is sufficient in the numerical result. We want to stress that one is limited to the fine resolution and a large CPU time if considering schemes such as ARS(2,3,3) or CGGS3.

5.3.2 Study of the order of accuracy

We now focus on the study of the order of accuracy of the schemes under consideration using the smooth solution (5.1).

The IMEX2, TVD2, MOOD2 and MOOD2(3) schemes. In Figure 8, we display the convergence curves for the four schemes, for $\varepsilon = 1$ (left panel) and $\varepsilon = 10^{-3}$ (right panel). As expected, we observe that the IMEX1 and TVD2 schemes are both first-order accurate, with the TVD2 scheme being more precise than the IMEX1 scheme. In addition, the MOOD2, MOOD2(3) and IMEX2 schemes are second-order accurate. Note that the error produced by the MOOD2(3) scheme is slightly larger than the one coming from the other two schemes. This is due to the less restrictive CFL condition of the MOOD2(3) scheme: were it lowered to match the one of the MOOD2 scheme, the errors would be comparable. This means that, in this context of a smooth solution, the MOOD correction allows us to get a second-order accurate scheme that also respects the maximum principle.

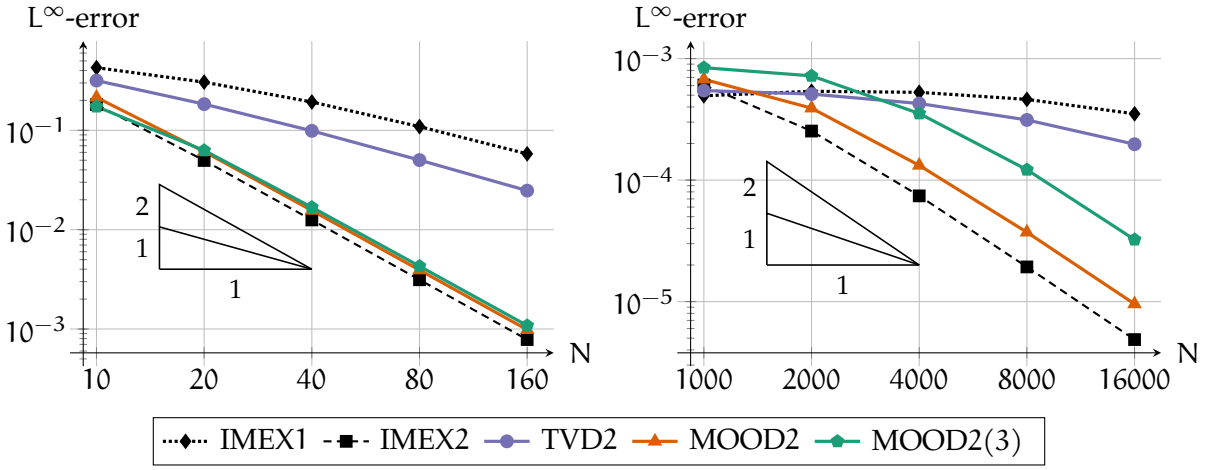

Figure 8: Error lines in L^∞ norm for the smooth solution (5.1) using the IMEX1, IMEX2, TVD2, MOOD2 and MOOD2(3) schemes. Left panel: $\varepsilon = 1$; right panel: $\varepsilon = 10^{-3}$.

The IMEX3, TVD3, MOOD3 and MOOD3(4) schemes. Now, let us consider the third-order IMEX3 scheme and the two schemes derived from this one. In Figure 9, we display the error to the exact solution, for $\varepsilon = 1$ in the left panel and $\varepsilon = 10^{-3}$ in the right panel. For $\varepsilon = 1$, we observe, as expected, that the TVD3 scheme is first-order accurate but more precise than the IMEX1 scheme, while the other three schemes are third-order accurate. For $\varepsilon = 10^{-3}$, we note that the error produced by the IMEX3 scheme starts to decrease slower than third-order when N becomes large. This is due to the instability of this IMEX3 scheme, and this problem is not experienced by the L-stable ARS(2,3,3) scheme. Due to these instabilities, the solution of the MOOD3 scheme is degraded since the MOOD algorithm switches more often to the TVD3 scheme than in the previous second-order case. Although its results are better, similar conclusions apply to the MOOD3(4) scheme.

5.3.3 Approximation of a discontinuous solution

Now, we study the numerical approximation of the discontinuous solution (5.2). Like in the previous Section, we first study the IMEX2, TVD2, MOOD2 and MOOD2(3) schemes, before moving on to the IMEX3, TVD3, MOOD3 and MOOD3(4) schemes. Lastly, we perform an experiment to show that the BDF2 and BDF3 discretizations alone violate the maximum principle.

Here, to compute the order of accuracy of the scheme, we no longer focus on the L^∞ norm, which is not suited to the computation of an error between a discontinuous solution and its diffusive

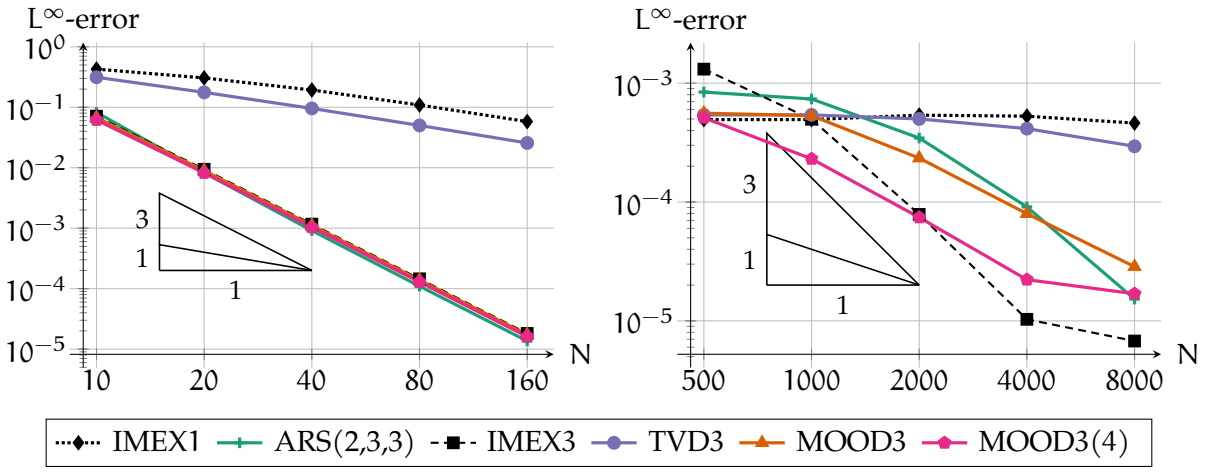

Figure 9: Error lines in L^∞ norm for the smooth solution (5.1) using the IMEX1, ARS(2,3,3), IMEX3, TVD3, MOOD3 and MOOD3(4) schemes. Left panel: $\varepsilon = 1$; right panel: $\varepsilon = 10^{-3}$.

approximation. Instead, we turn to the L^1 norm, defined by

$$\|w^n\|_1 = \frac{1}{\Delta x} \sum_j |w_j^n|.$$

However, the above norm only measures the average deviation between the exact solution and the numerical approximation. Here, since we seek a measure of the maximum principle violation of the IMEXp scheme, we instead consider the following modification of the L^1 norm:

$$\|w^n\|_{L^1_o} = \frac{1}{\Delta x} \sum_j \left(|w_j^n| + \max_{m \leqslant n} \left[\left(\max_j w_j^m - \min_j w_j^m \right) - \left(\max_j w_j^0 - \min_j w_j^0 \right) \right] \right).$$

This quantity, although it does not satisfy the triangle inequality property of a norm, as it is in fact a quasinorm, allows us to add the impact of overshoots and undershoots to the usual measure of the average deviation between the solution and its approximation. Since the TVDp and MOODp methods are built to avoid such over- and undershoots, this additional term will vanish with these methods.

The IMEX2, TVD2, MOOD2 and MOOD2(3) schemes. In Figure 10, we display the results of the four schemes, and of the IMEX1 scheme for the sake of comparison, when approximating the discontinuous solution (5.2) (left panel: $\varepsilon = 1$, right panel: $\varepsilon = 10^{-3}$). In both cases, we observe that the IMEX2 scheme violates the maximum principle, while it is satisfied by the other four schemes. We observe that both phase and amplitude errors are present. Like before, taking the largest possible CFL number condition for the MOOD2(3) worsens its precision, but this behaviour merely highlights the flexibility of large time step schemes, as taking a more restrictive CFL condition would increase the precision at the cost of CPU time.

In Figure 11, we display the error lines in L^1 norm (left panels) and L^1_o quasinorm (right panels), for $\varepsilon = 1$ (top panels) and $\varepsilon = 10^{-3}$ (bottom panels). First, we observe that the theoretical order of convergence is not reached. At most, the schemes are order $\frac{1}{2}$. This is due to the fact that we approximate a discontinuous solution, where the numerical diffusion of the schemes considerably worsen the order of convergence, see for instance [24], Chapter 11. Second, as expected, the L^1-error of the IMEX2 scheme is lower than the one of the other schemes. Also, when taking the over- and

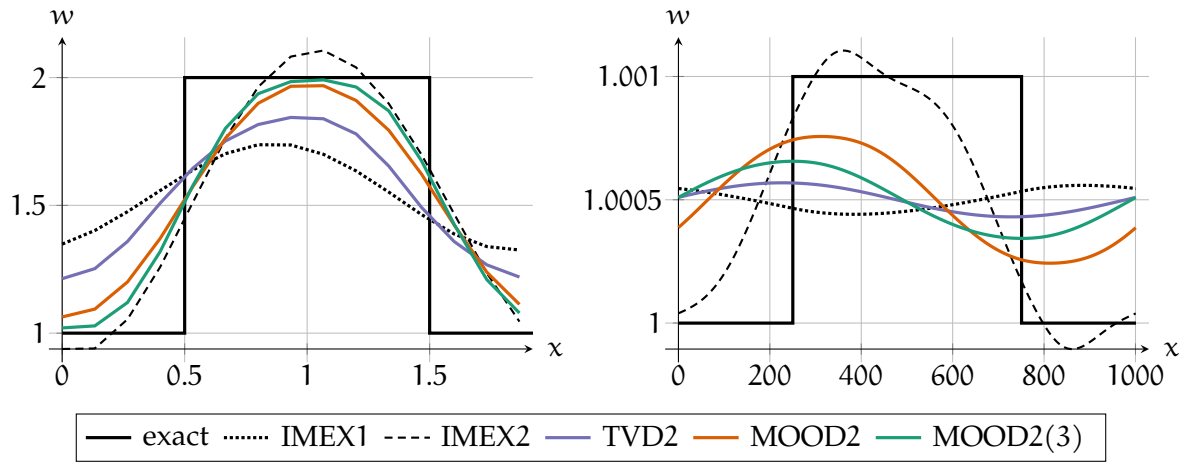

Figure 10: Approximation of the discontinuous solution (5.2) at time t_{end} using the IMEX1, IMEX2, TVD2, MOOD2 and MOOD2(3) schemes. Left panel: $\varepsilon = 1$ and $N = 15$; right panel: $\varepsilon = 10^{-3}$ and $N = 2000$.

undershoots into account thanks to the L^1_o quasinorm, we observe that the L^1_o quasinorm of the error produced by the IMEX1, TVD2, MOOD2 and MOOD2(3) schemes is the same as their L^1 norm. This was to be expected since no over- or undershoots are produced by these schemes. However, when looking at the L^1_o quasinorm of the error of the IMEX2 scheme, we observe that it stays roughly constant as N grows larger. This means that the improvement in L^1 norm, since the numerical solution is overall closer to the exact solution, is almost exactly compensated by an increase of the over- and undershoot magnitude. Therefore, even taking large N is not enough to ensure a good approximation of the exact discontinuous solution by the IMEX2 scheme.

The IMEX3, TVD3, MOOD3 and MOOD3(4) schemes. Now, we turn to Figure 12, where we have displayed the numerical approximation of the discontinuous solution by the IMEX1, ARS(2,3,3), IMEX3, TVD3, MOOD3 and MOOD3(4) schemes, for $\varepsilon = 1$ in the left panel and $\varepsilon = 10^{-3}$ in the right panel. Once again, we note that the pure IMEX high-order schemes are oscillatory and violate the maximum principle, while the other four schemes are in-bounds. A notable remark concerns the IMEX3 scheme when $\varepsilon = 10^{-3}$, in the right panel depicted by the dashed line. In this case, the scheme is so unstable that the numerical solution is unrecognisable. The MOOD3 scheme corrects this shortcoming. Furthermore, the MOOD3(4) scheme is based on a more stable third-order scheme, which ensures a better approximation of the exact solution.

In Figure 13, we report the error produced by the six schemes, in the L^1 norm in the left panels and in the L^1_o quasinorm in the right panels, for $\varepsilon = 1$ in the top panels and $\varepsilon = 10^{-3}$, except for the IMEX3 scheme, whose error would explode in the bottom panels. Like in the case of the second-order schemes, we observe that the theoretical order of convergence is not reached, and that the schemes are accurate up to order $\frac{1}{2}$ for the IMEX1, TVD3, MOOD3 and MOOD3(4) schemes, and up to order $\frac{3}{4}$ for the ARS(2,3,3) and IMEX3 schemes. In addition, once again, the L^1_o quasinorm for the ARS(2,3,3) and IMEX3 schemes stays roughly constant as N increases, which means that the L^1-error improvement is compensated by an increase in the over- and undershoot amplitude.

5.4 Application to the isentropic Euler equations

In these last numerical experiments, we consider an application of the IMEXp, TVDp and MOODp schemes to the isentropic Euler equations. This system models a compressible fluid flow. It is governed,

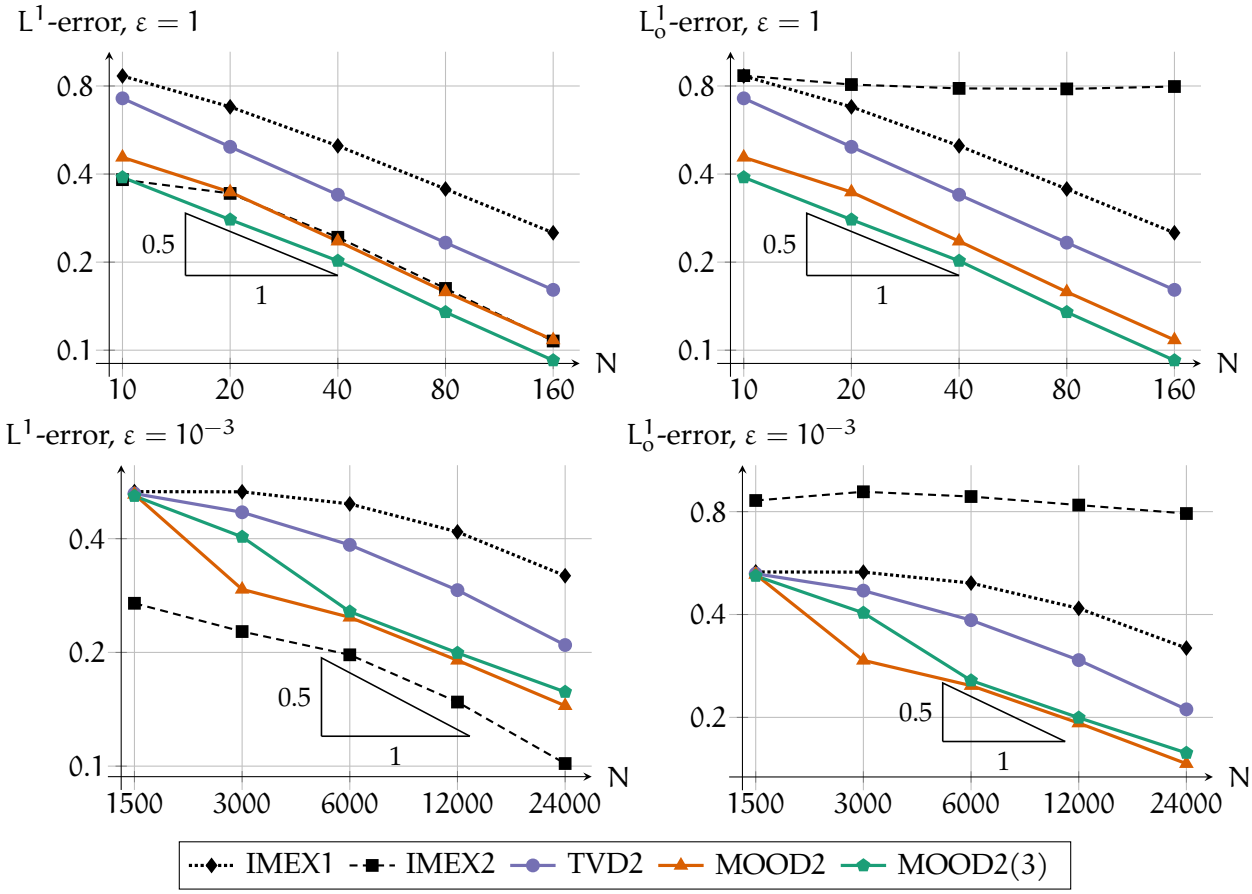

Figure 11: Error lines in L^1 norm (left panels) and L^1_o quasinorm (right panels) for the discontinuous solution (5.2) using the IMEX1, IMEX2, TVD2, MOOD2 and MOOD2(3) schemes. Top panels: $\varepsilon = 1$; bottom panels: $\varepsilon = 10^{-3}$.

in one space dimension and after a suitable rescaling, by:

$$\begin{cases} \partial_t \rho + \partial_x(\rho u) = 0, \\ \partial_t(\rho u) + \partial_x(\rho u^2) + \dfrac{1}{M^2}\partial_x p(\rho) = 0, \end{cases} \tag{5.6}$$

where $\rho(x,t) > 0$ is the fluid density, $u(x,t)$ is its velocity. Assuming an ideal gas, the pressure is $p(\rho) = \rho^\gamma$, with $\gamma \geqslant 1$ the ratio of specific heats. Finally, M is the Mach number, which represents the ratio of material to acoustic velocity. The system (5.6) represents a compressible flow for $M > 0$, whereas the flow becomes incompressible when M tends to 0. For more information, see for instance [22, 26].

We are specifically concerned with asymptotic-preserving schemes, i.e. schemes that behave correctly in the low Mach number limit. Such schemes have been the focus of much work in the recent past, see for instance [8, 10, 4], but this list is far from being exhaustive. In this section, we focus on the asymptotic-preserving scheme derived in [10], whose semi-discretization in time reads:

$$\frac{\rho^{n+1} - \rho^n}{\Delta t} + \partial_x(\rho u)^{n+1} = 0, \tag{5.7a}$$

$$\frac{(\rho u)^{n+1} - (\rho u)^n}{\Delta t} + \partial_x(\rho u^2)^n + \frac{1}{M^2}\partial_x p(\rho^{n+1}) = 0. \tag{5.7b}$$

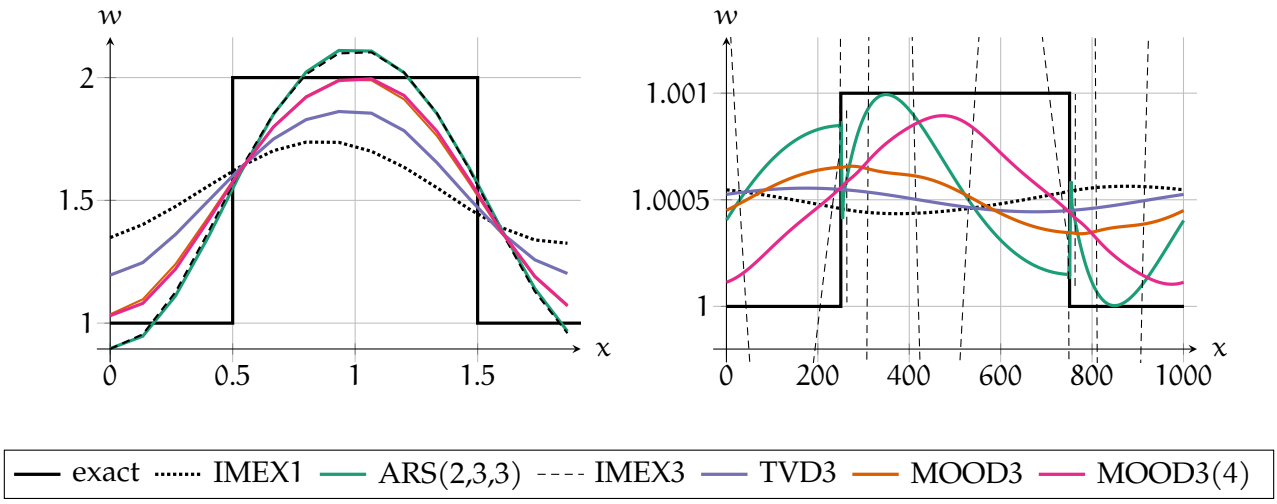

Figure 12: Approximation of the discontinuous solution (5.2) at time t_{end} using the IMEX1, ARS(2,3,3), IMEX3, TVD3 and MOOD3 schemes. Left panel: $\varepsilon = 1$ and $N = 15$; right panel: $\varepsilon = 10^{-3}$ and $N = 1500$. In the right panel, the errors produced by the IMEX3 scheme have destroyed the numerical approximation.

We recast (5.7) under the following condensed form:

$$\frac{W^{n+1} - W^n}{\Delta t} + \partial_x F_e(W^n) + \partial_x F_i(W^{n+1}) = 0, \tag{5.8}$$

where we have set

$$W = \begin{pmatrix} \rho \\ \rho u \end{pmatrix}; \quad F_e(W) = \begin{pmatrix} 0 \\ \rho u^2 \end{pmatrix}; \quad F_i(W) = \begin{pmatrix} \rho u \\ \frac{1}{M^2} p(\rho) \end{pmatrix}, \tag{5.9}$$

with F_e and F_i respectively being the explicit and implicit fluxes. Note that the seemingly coupled system (5.7) can be decoupled by inserting the value of $(\rho u)^{n+1}$ from (5.7b) into (5.7a). Also, we remark that, by adapting our TVD MOOD time integration to the semi-discretized scheme (5.8), we recover an asymptotic-preserving scheme, analogously to [10, 9].

There is a natural correspondence between the Euler equations (5.8) and the toy problem (2.1). Indeed, $F_e(W)$ corresponds to $c_m w$ and $F_i(W)$ corresponds to $\frac{c_a}{\varepsilon} w$, with ε representing the square Mach number M^2. Therefore, once a suitable space discretisation is chosen for (5.8), applying the IMEXp and TVDp schemes is straightforward in either the finite difference or the finite volume framework. For the TVDp scheme, the same value of θ that was derived for the toy problem is directly used for the Euler system. However, adapting the MOODp scheme requires the introduction of a new detection criterion.

Indeed, since the Euler equations (5.6) form a hyperbolic system of conservation laws, the basic MOOD criterion (DMP) on the L^∞ norm of the unknown w from Section 5.1 is no longer valid. Instead, we follow [9] and use the Riemann invariants to detect oscillations, since we know from [35] that at least one of the Riemann invariants satisfies a maximum principle in a Riemann problem. The Riemann invariants are given by:

$$\Phi_\pm(W) = u \mp \frac{1}{M} \frac{2}{\gamma - 1} \sqrt{\gamma \rho^{\gamma - 1}}.$$

We thus adapt the MOOD algorithm (Algorithm 7) for the Euler equations as follows.

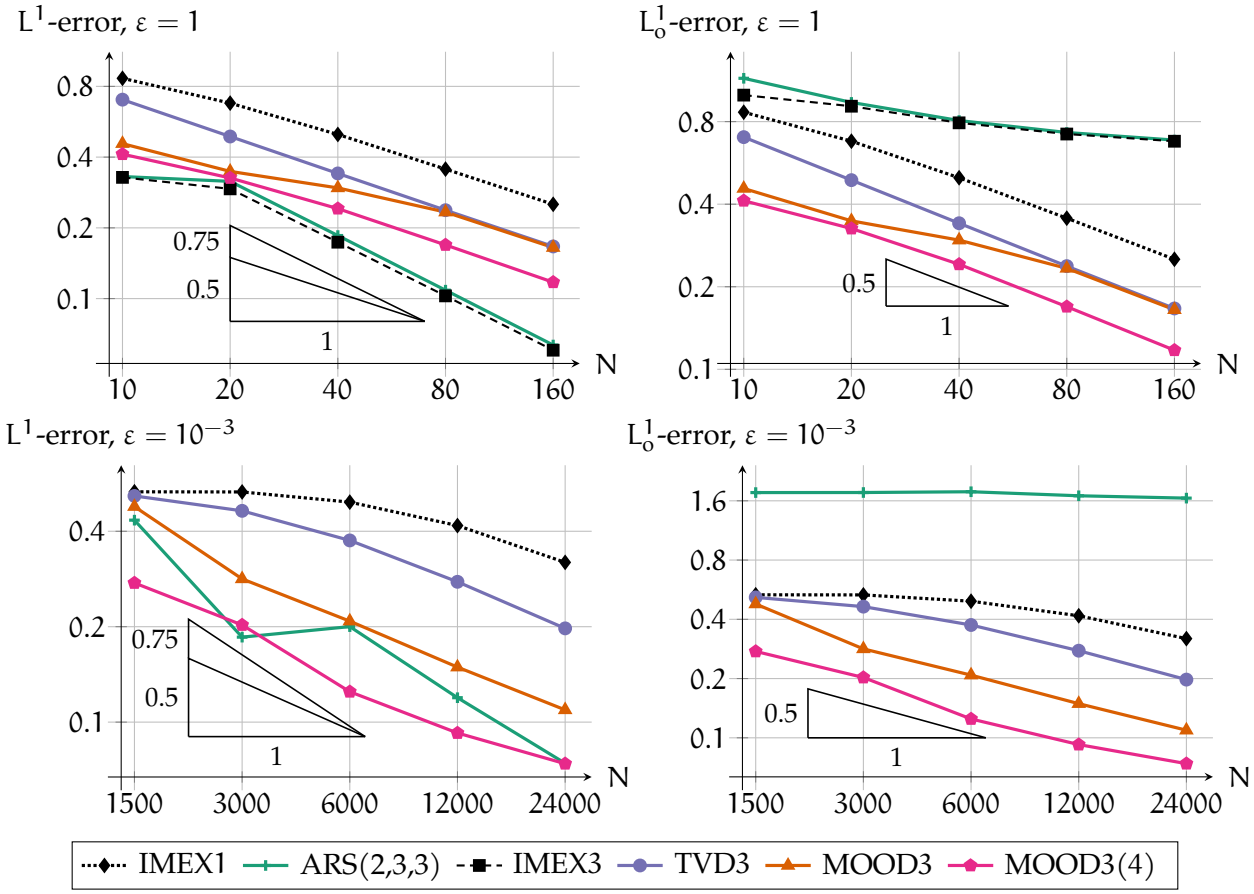

Figure 13: Error lines in L^1 norm (left panels) and L^1_o quasinorm (right panels) for the discontinuous solution (5.2) using the IMEX1, ARS(2,3,3), IMEX3, TVD3, MOOD3 and MOOD3(4) schemes. Top panels: $\varepsilon = 1$; bottom panels: $\varepsilon = 10^{-3}$. For $\varepsilon = 10^{-3}$, the IMEX3 error is so large that the error lines are not displayed (see Figure 12, right panel).

Algorithm 9 (MOODp scheme for the isentropic Euler equations)**.** *Define the initial detection criterion* $\mathcal{E}^0_\pm = \|\Phi_\pm(W^0)\|_\infty$. *Equipped with the stable TVD*p *scheme, the MOOD*p *scheme consists in applying the following procedure at each time step:*

1. *Compute a candidate numerical solution* W_c^{n+1} *with the IMEX*p *scheme.*
2. *Detect whether the discrete maximum principle is satisfied by the Riemann invariants:*

$$\|\Phi_\pm(W_c^{n+1})\|_\infty \leqslant \mathcal{E}^n_\pm. \tag{DMP}$$

(3a) *If* (DMP) *holds, then set the numerical solution* W^{n+1} *equal to the candidate solution* W_c^{n+1}.

(3b) *Otherwise, compute the numerical solution* W^{n+1} *with the TVD*p *scheme.*

(4) *Update the detection criterion with* $\mathcal{E}^{n+1}_\pm = \xi\|\Phi_\pm(W^{n+1})\|_\infty + (1-\xi)\mathcal{E}^n_\pm$.

Remark that, compared to Algorithm 7, the detection criterion is relaxed with a convex combination of parameter $\xi \in [0,1]$ between the current solution and the previous time steps. This allows a finer control over how oscillatory we allow the MOOD solution to be. Indeed, the closer ξ is to 0, the

less permissive the MOOD procedure will be of small oscillations. Unless otherwise specified, we take $\xi = \frac{1}{2}$ in the following experiments.

Regarding the space discretisation, we focus on the finite volume scheme proposed in [10], which has been built to ensure the L^∞ stability on a linearized version of (5.6), at the cost of some extra numerical viscosity. For the sake of conciseness, we do not rewrite this space discretisation here; the reader is referred to [10, 9] for more information. This space-time discretisation allows us to take a time step constrained by

$$\Delta t \leqslant C \frac{\Delta x}{2 \max_j u_j},$$

which does not depend on the Mach number M, unlike in the case of classical explicit schemes.

Equipped with the IMEXp, TVDp and MOODp schemes for the isentropic Euler equations, we now apply them, first to a smooth solution to test the order of accuracy of the schemes, and then to a discontinuous solution to test the TVD property. For the sake of conciseness, we only display the results of the IMEX3(4), TVD3(4) and MOOD3(4) schemes, but the conclusions hold for the other schemes tested in Section 5.3. In addition, we compare our results to those of the TVD2 and MOOD2 schemes, which were introduced and used in [9].

In the following experiments, we prescribe homogeneous Neumann boundary conditions on the space domain $(0, 1)$.

5.4.1 Order of accuracy

We have shown in Section 5.3.2 that the schemes, applied to the model problem, exhibit the expected order of accuracy. We now compute the order of accuracy of our schemes in this context of the isentropic Euler equations. To that end, we consider the procedure proposed in [37] and used in [9]. To obtain an exact solution, we choose the well-prepared initial condition

$$\rho(0, x) = 1 - \frac{M^2}{2}\omega\left(4\left(x - \frac{1}{2}\right)\right) \quad \text{and} \quad u(0, x) = 1 + \frac{M^2}{2}\omega\left(4\left(x - \frac{1}{2}\right)\right),$$

where ω is a classical compactly supported smooth function, given by

$$\omega(z) = \begin{cases} \exp\left(1 - \dfrac{1}{1 - z^2}\right) & \text{if } z < 1, \\ 0 & \text{otherwise.} \end{cases}$$

To compute the exact solution at time t, one needs to consider $\gamma = 3$ and follow the Riemann invariants. The procedure is explained in detail in the two references above, and we do not repeat it here.

In Figures 14 and 15, we report at the final time $t_{end} = 0.03M$ the L^∞ errors produced on ρ and ρu by the four schemes for $M = 1$ and $M = 10^{-2}$ respectively. The schemes behave as expected, i.e. as they did in the toy problem case.

5.4.2 Approximation of a Riemann problem

Next, we apply the schemes to the approximation of a discontinuous solution. In this context of the isentropic Euler equations, we propose a simulation of the following Riemann problem with well-prepared initial data:

$$\begin{cases} \rho(x, 0) = \begin{cases} 1 + M^2 & \text{if } x < 0.5, \\ 1 & \text{otherwise,} \end{cases} \\ (\rho u)(x, 0) = 0.25. \end{cases} \tag{5.10}$$

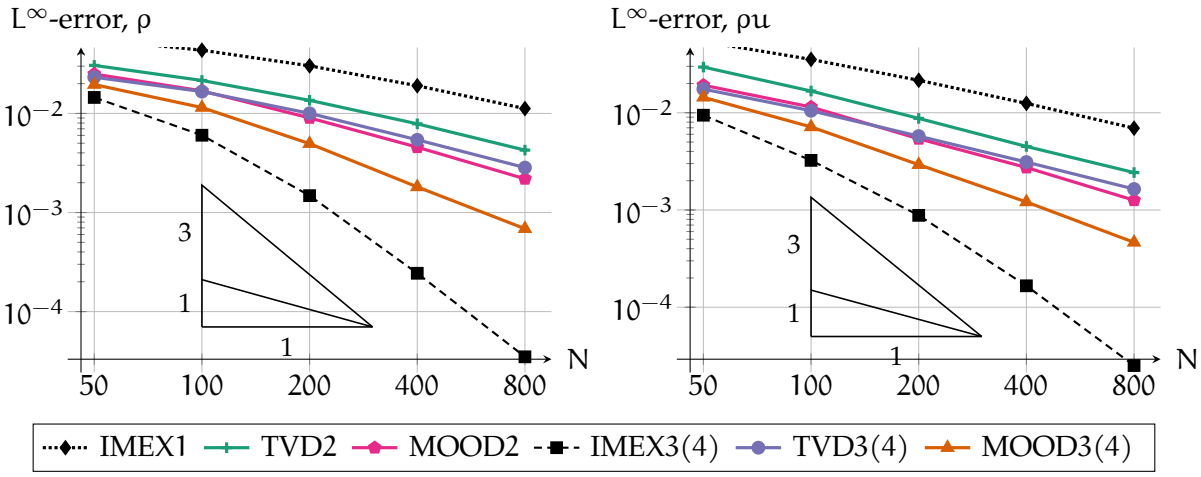

Figure 14: Error lines in L^∞ norm for the smooth solution described in Section 5.4.1, with $M = 1$ and using the IMEX1, TVD2, MOOD2, IMEX3(4), TVD3(4), and MOOD3(4) schemes. Left panel: density ρ; right panel: momentum ρu.

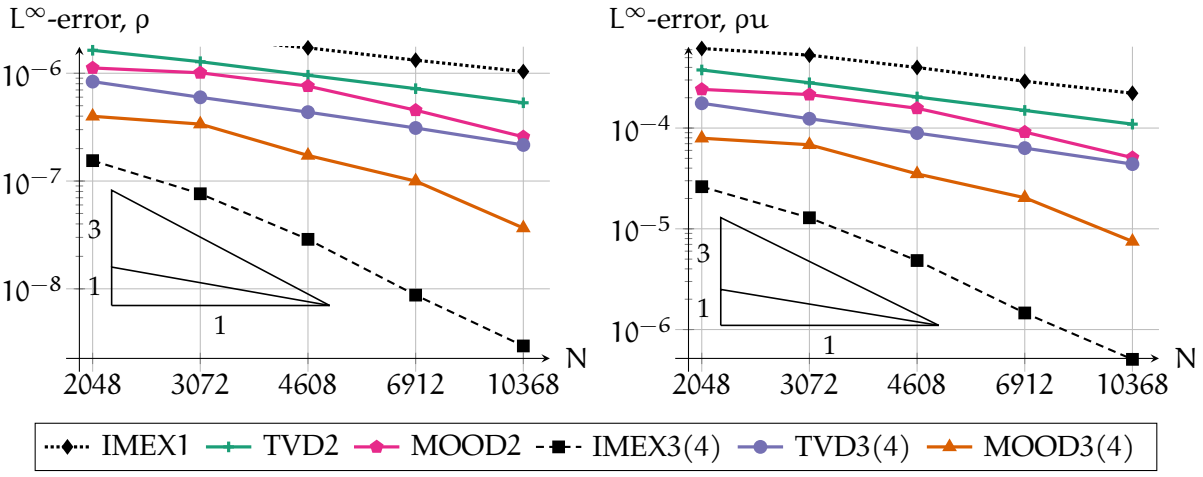

Figure 15: Error lines in L^∞ norm for the smooth solution described in Section 5.4.1, with $M = 10^{-2}$ and using the IMEX1, TVD2, MOOD2, IMEX3(4), TVD3(4), and MOOD3(4) schemes. Left panel: density ρ; right panel: momentum ρu.

We take $\gamma = 1.4$ and we compute the solution until the final time $t_{end} = 0.15M$. Also, to eliminate more oscillations, we take $\xi = \frac{1}{20}$ for $M = 1$. The approximations are depicted in Figures 16 and 17 for, respectively, $M = 1$ with $N = 50$, and $M = 10^{-2}$ with $N = 2500$. We have elected to represent only the MOOD2 and MOOD3(4) results for the sake of clarity in the figures.

Similar conclusions as in Section 5.3.3 are drawn from this experiment. Note that the IMEX3(4) result is not displayed in Figure 17, since it is too oscillatory. In addition, for $M = 1$ and $M = 10^{-2}$, the MOOD procedure is activated respectively on 51% and 50% of time iterations. This observation is explained by the fact that the IMEX3(4) scheme is quite oscillatory, and its oscillations have to be countered by the MOOD procedure. The share of iterations where the MOOD procedure was activated could be lowered by basing the MOOD procedure on a less oscillatory third-order IMEX scheme than the IMEX3(4) scheme.

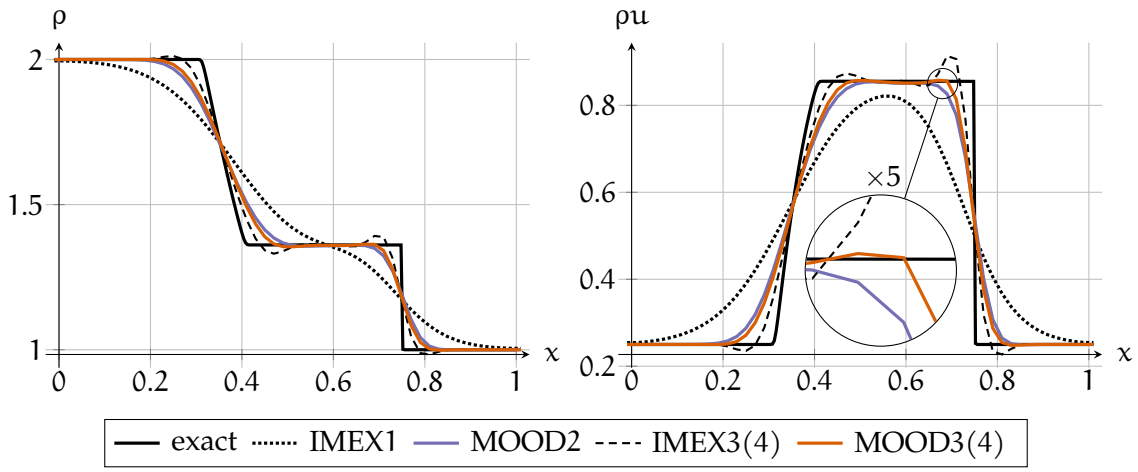

Figure 16: Approximation of the solution to the Riemann problem (5.10) at time t_{end} with $M = 1$, $N = 50$, and using the IMEX1, MOOD2, IMEX3(4) and MOOD3(4) schemes. Left panel: density ρ; right panel: momentum ρu.

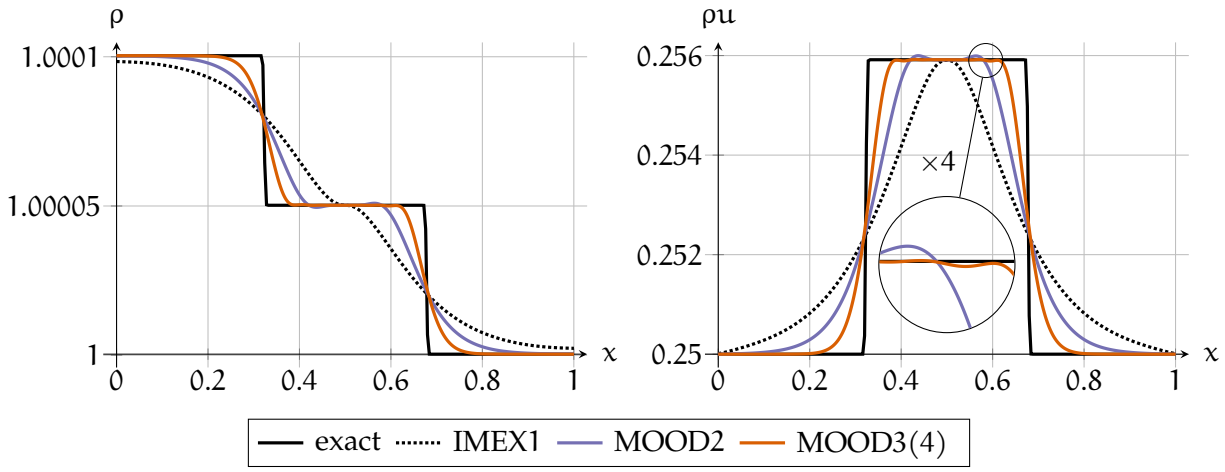

Figure 17: Approximation of the solution to the Riemann problem (5.10) at time t_{end} with $M = 10^{-2}$, $N = 2500$, and using the IMEX1, MOOD2, and MOOD3(4) schemes. Left panel: density ρ; right panel: momentum ρu.

Also, remark that our new MOOD3(4) scheme is more accurate than the MOOD2 scheme from [9], as evidenced by the two zooms on the momentum shock wave, especially for $M = 10^{-2}$, displayed in Figure 17, where we observe a significant improvement in the shock profile, as well as in the reduction of oscillations.

Furthermore, as expected, these schemes allow us to use a time step much larger than classical explicit schemes when the Mach number is low. Indeed, for $M = 1$, we get a time step of the same order as the classical one, whereas in the low Mach number regime where $M = 10^{-2}$, we can use the time step $\Delta t \simeq 1.9 \times 10^{-4}$, which is about $10^2 = \frac{1}{M}$ times larger than the classical explicit scheme with $\Delta t \simeq 8.3 \times 10^{-7}$.

6 Conclusions and future work

We have presented a new approach to construct high-resolution TVD IMEX-RK schemes for multi-scale equations which are computationally efficient. Motivated by the order barrier for unconditionally stable implicit and conditionally stable IMEX Runge schemes that only have a CFL restriction depending on the explicitly treated terms, we introduced a new class of TVD schemes consisting of a convex combination with a first-order TVD IMEX scheme and a high-order IMEX RK scheme. Even if the TVD property is not satisfied by the high-order schemes for regions with discontinuities, as displayed in Figures 10 and 12, it is verified in smooth enough regions. To recover accuracy in such regions, we have combined our first-order TVD schemes with a MOOD procedure. In Figures 8, 9, 11 and 13, we saw that our schemes perform well when compared with schemes from the literature. Due to having the option of taking large time steps, the schemes are suitable for applications where the focus of the numerical solution is on the usually slow explicitly treated dynamics, with the added flexibility of ensuring the L^∞ stability for various CFL restrictions, as seen in Table 2. Finally, we successfully applied the schemes to the isentropic Euler equations, which confirmed that our approach improved on previous results from [9], especially for small Mach numbers.

In this paper, the focus was to give a theoretical justification of our TVD approach by means of studying a one dimensional linear scalar equation. The application to the isentropic Euler system confirmed the validity of our approach when applied to non-linear systems, and showed an increase in precision and stability compared to the existing scheme in [9]. However, here, we did not consider the problematic of scale dependent diffusion and multi-dimensional problems, which will be addressed in future work.

Acknowledgements : V. Michel-Dansac extends his thanks to the Service Hydrographique et Océanographique de la Marine (SHOM) for financial support. A. Thomann acknowledges the support of the INDAM-DP-COFUND-2015, grant number 713485. This work was started during the SHARK-FV conference (Sharing Higher-order Advanced Research Know-how on Finite Volume http://www.SHARK-FV.eu/) held in 2019. The authors would also like to thank Gabriella Puppo for fruitful discussions and comments.

A Proving the incompatibility of high-order TVD IMEX schemes with large time steps

We write the IMEX update (2.3) as a convex combination of forward and backward Euler steps, in the notation of [34, 15], as

$$w^{n+1} = (1-h)\sum_{k=0}^{i-1}\left(\alpha_{ik}w^{(k)} + \Delta t\frac{\tilde{\beta}_{ik}}{1-h}c_m w_x^{(k)}\right) + h\left(\sum_{k=0}^{i-1}\alpha_{ik}w^{(k)} + \Delta t\frac{\beta_i}{h}\frac{c_a}{\varepsilon}w_x^{(i)}\right), \tag{A.1}$$

where $h \in (0,1)$ and the weights $\alpha_{ik} \geqslant 0$ fulfilling $\sum \alpha_{ik} = 1$. We assume $\beta_i > 0$ and $\tilde{\beta}_{ik} \geqslant 0$ without loss of generality. Indeed, negative β_i or $\tilde{\beta}_{ik}$ could still yield a TVD scheme, by changing the upwinding direction in the discretisation of the derivatives w_x, as suggested in [15]. For simplicity, we also assume without loss of generality, in accordance with [15], that the non-diagonal entries of the implicit Butcher tableau are zero. We immediately see from (A.1) that the explicit part is a convex combination of TVD forward Euler (fE) steps, and is thus TVD under the CFL restriction

$$\Delta t \leqslant (1-h)\,\min\left(\frac{\alpha_{ik}}{\tilde{\beta}_{ik}}\right)\Delta t_{fE} \tag{A.2}$$

where the minimum is set to infinity if $\tilde{\beta}_{ik} = 0$. In [15], it was shown that a necessary condition for the TVD property is that all weights α_{ik} have to be positive while satisfying the order conditions. Unfortunately, the authors prove in the same work that this is impossible for implicit schemes with order $p \geqslant 2$ which do not require a CFL condition. Analogously, we show that this result still holds for the IMEX update (A.1) under the scale-independent CFL condition (A.2).

Proposition 10. *For an IMEX-RK update* (A.1)*, under a scale-independent CFL condition* (A.2) *of order* $p \geqslant 2$*, there is at least one negative* α_{ik}*.*

Proof. The order conditions for high-order RK schemes are included in the compatibility conditions for higher order IMEX-RK schemes. The second-order conditions read for the implicit part of (A.1) as

$$\sum_{k=0}^{i-1} \alpha_{ik} = 1, \quad X_s = h, \quad Y_s = \frac{1}{2}h^2, \tag{A.3}$$

where $h \in (0, 1)$, and where X_s, Y_s are defined recursively by

$$X_1 = \beta_1, \quad Y_1 = \beta_1^2, \quad X_s = \beta_s + \sum_{i=1}^{s-1} \alpha_{si} X_i, \quad Y_s = \beta_s X_s + \sum_{i=1}^{s-1} \alpha_{si} Y_i.$$

Following the proof in [15], we show now that, if $\alpha_{ik} \geqslant 0$ for all i, k, then we get

$$hX_s - Y_s < \frac{1}{2}h^2,$$

which contradicts (A.3). This contradiction is shown by using the formula

$$(1-\zeta)hX_s - Y_s \leqslant \tau_s(1-\zeta)^2$$

with arbitrary $\zeta \in \mathbb{R}$ and

$$0 < \tau_1 = \frac{1}{4}h^2, \quad \tau_s = \frac{h^4}{4(h^2 - \tau_{s-1})}. \tag{A.4}$$

This estimate is shown by induction following the steps given in [15]. From (A.4), we find

$$0 < \tau_1 = \frac{1}{4}h^2 < \ldots < \tau_s < \frac{1}{2}h^2$$

which completes the proof. □

B On the incompatibility of BDF with TVD

Using for instance a second order Backward-Differencing-Formula (BDF), see [1], to approximate the implicit derivative, leads to

$$\frac{\partial w(x,t)}{\partial x} \approx \frac{1}{\Delta x}\left(3w_j - 4w_{j-1} + w_{j-2}\right), \tag{B.1}$$

while the third-order BDF approximation is given by

$$\frac{\partial w(x,t)}{\partial x} \approx \frac{1}{\Delta x}\left(\frac{11}{6}w_j - 3w_{j-1} + \frac{3}{2}w_{j-2} - \frac{1}{3}w_{j-3}\right). \tag{B.2}$$

Using the second-order BDF (B.1) in the first step of the scheme (3.2), we get

$$w_j^{(2)} + \mu_\varepsilon \frac{a_{22}}{2}\left(3w_j^{(2)} - 4w_{j-1}^{(2)} + w_{j-2}^{(2)}\right) = w_j^n - \lambda a_{22}\Delta_j^n.$$

Following the proof from Lemma 1, we have

$$\begin{aligned}\|w^n\|_\infty &\geqslant \max_j \left|\left(1 + \mu_\varepsilon \frac{3a_{22}}{2}\right) w_j^{(2)} - \mu_\varepsilon \frac{a_{22}}{2}\left(4w_{j-1}^{(2)} - w_{j-2}^{(2)}\right)\right| \\ &\geqslant \left(1 + \mu_\varepsilon \frac{3a_{22}}{2}\right)\|w^{(2)}\|_\infty - \mu_\varepsilon \frac{a_{22}}{2}\max_j \left|4w_{j-1}^{(2)} - w_{j-2}^{(2)}\right|\end{aligned}$$

To complete this step we need

$$\max_j \left|4w_{j-1}^{(2)} - w_{j-2}^{(2)}\right| \leqslant 4\|w^{(2)}\|_\infty - \|w^{(2)}\|_\infty \tag{B.3}$$

which is a contradiction to the inverse triangular equation. Therefore using a second-order BDF does not lead to a TVD scheme. We can even extend this observation to a BDF of general order. As it is derived to match the Taylor series expansion up to an order p, its general form has alternating signs, and it can be written using $p+1$ coefficients $\kappa_i \geqslant 0$, $i = 0, \ldots, p$, as in [1]

$$\frac{\partial w(x,t)}{\partial x} \approx \kappa_0 w_j - \kappa_1 w_{j-1} + \kappa_2 w_{j-2} - \ldots + \kappa_p w_{j-p} \tag{B.4}$$

for an approximation of order p, where we have taken an even p for the moment. We use the BDF described by (B.4) for the approximation of the implicit space derivative, and we find in the estimate for the L^∞ stability:

$$\begin{aligned}\|w^n\|_\infty &\geqslant \max_j \left|(1 + \mu_\varepsilon a_{22}\kappa_0) w_j^{(2)} - \mu_\varepsilon a_{22}\left(\kappa_1 w_{j-1}^{(2)} - \kappa_2 w_{j-2}^{(2)} + \kappa_3 w_{j-3}^{(2)} - \ldots - \kappa_m w_{j-m}^{(2)}\right)\right| \\ &\geqslant (1 + \mu_\varepsilon a_{22}\kappa_0)\|w^{(2)}\|_\infty - \mu_\varepsilon a_{22}\max_j \left|\kappa_1 w_{j-1}^{(2)} - \kappa_2 w_{j-2}^{(2)} + \kappa_3 w_{j-3}^{(2)} - \ldots - \kappa_m w_{j-m}^{(2)}\right| \\ &\geqslant (1 + \mu_\varepsilon a_{22}\kappa_1)\|w^{(2)}\|_\infty - \mu_\varepsilon a_{22}\max_j \left|\kappa_1 w_{j-1}^{(2)} - \kappa_2 w_{j-2}^{(2)}\right| - \ldots \\ &\qquad - \mu_\varepsilon a_{22}\max_j \left|\kappa_{p-1} w_{j-p-1}^{(2)} - \kappa_p w_{j-p}^{(2)}\right|.\end{aligned}$$

Analogously to (B.3), to achieve the right estimate, the inverse triangular inequality would be violated. The case of an odd p also fails.

C On non-CK IMEX schemes

Consider the following Butcher tableaux, defining an IMEX scheme in non-CK, non-ARS form:

$$\text{explicit:}\quad \begin{array}{c|cccc} 0 & 0 & 0 & \cdots & 0 \\ \tilde{c}_2 & \tilde{a}_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \tilde{c}_s & \tilde{a}_{s1} & \cdots & \tilde{a}_{s,s-1} & 0 \\ \hline & \tilde{b}_1 & \cdots & \tilde{b}_{s-1} & \tilde{b}_s \end{array} \qquad \text{implicit:}\quad \begin{array}{c|cccc} c_1 & a_{11} & 0 & \cdots & 0 \\ c_2 & a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}. \tag{C.1}$$

We derive stability conditions analogous to Theorem 4 for this case where the first column of the implicit tableau is non-zero. After lengthy computations, we get the following result:

Theorem 11. *Let $\tilde{A}, A \in \mathbb{R}^{s\times s}$, $\tilde{b}, b, \tilde{c}, c \in \mathbb{R}^s$ define two Butcher tableaux (C.1) fulfilling (2.6) and the p-th order compatibility conditions. Let $\tilde{b}$ and b coincide with the last rows of $\tilde{A}$ and A respectively. For $k = 1, \dots, s$ and $l = 1, \dots, k-1$, we define*

$$\mathcal{A}_k = \theta_k a_{kk} + (1-\theta_k)c_k, \quad \tilde{\mathcal{A}}_k = (1-\theta_k)\tilde{c}_k, \quad \mathcal{B}_{kl} = \frac{\theta_k a_{kl}}{\mathcal{A}_l}, \quad \tilde{\mathcal{B}}_{kl} = \theta_k \tilde{a}_{kl}.$$

In addition, we recursively define the following expressions:

$$\tilde{\mathcal{C}}_k = \tilde{\mathcal{A}}_k - \sum_{l=2}^{k-1} \mathcal{B}_{kl}\tilde{\mathcal{C}}_l, \qquad \tilde{\mathcal{D}}_{kl} = \tilde{\mathcal{B}}_{kl} - \sum_{r=l+1}^{k-1} \mathcal{B}_{kr}\tilde{\mathcal{D}}_{rl},$$
$$\mathcal{C}_k = 1 - \sum_{l=1}^{k-1} \mathcal{B}_{kl}\mathcal{C}_l, \qquad \mathcal{D}_{kl} = \mathcal{B}_{kl} - \sum_{r=l+1}^{k-1} \mathcal{B}_{kr}\mathcal{D}_{rl}.$$

Then, under the following restrictions for $k = 1, \dots, s$ and $l = 1, \dots, k-1$,

$$\mathcal{A}_k > 0, \qquad 0 \leqslant \lambda\tilde{\mathcal{C}}_k \leqslant \mathcal{C}_k, \qquad 0 \leqslant \lambda\tilde{\mathcal{D}}_{k,l} \leqslant \mathcal{D}_{k,l},$$

the scheme consisting in the convex combination based on the Butcher tableaux (C.1), combined with a TVD limiter, is L^∞ stable and TVD under a CFL condition determined by $\lambda \geqslant 0$ where λ does not depend on ε.

When performing numerical experiments, we observe that the results of schemes derived under the conditions of Theorem 11 are not as compelling as results of schemes obeying Theorem 4. Therefore, we do not include such schemes in the numerical experiments, but we still state Theorem 11 for the sake of completeness.

D TVD2(3)

For the TVD2(3) scheme, define $a_{32} = 0.3280595784620364$ and $a_{33} = 0.3386070882046304$. Then, the Butcher tableaux are given by:

explicit:
$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 \\ 2/3 & a_{32} & a_{33} & 0 \\ \hline & 0 & 1/2 & 1/2 \end{array},$$

implicit:
$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 \\ 2/3 & 0 & a_{32} & a_{33} \\ \hline & 0 & 1/2 & 1/2 \end{array}.$$

E TVD3(4)

For the TVD3(4) scheme, the explicit Butcher tableau is given by:

0	0	0	0	0
0.2049503677289891	0.2049503677289891	0	0	0
0.4173127343286904	0.2123925641886599	0.2049201701400305	0	0
0.9048203025659662	−0.4501877125339555	0.3955748607480934	0.9594331543518283	0
	0	0.3354718384287510	0.3487815573407456	0.3157466042305059

while the implicit Butcher tableau is given as follows:

0	0	0	0	0
0.2049503677289891	0	0.2049503677289891	0	0
0.4173127343286904	0	0.2040104873103189	0.2133022470183705	0
0.9048203025659662	0	0.3991926529002874	0.4115004113464103	0.0941272383192684
	0	0.3354718384287510	0.3487815573407456	0.3157466042305059

.

References

[1] U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. SIAM: Society for Industrial and Applied Mathematics, 1998.

[2] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri. Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations. *Appl. Numer. Math.*, 25(2-3):151–167, 1997. Special issue on time integration (Amsterdam, 1996).

[3] G. Bispen, K. R. Arun, M. Lukáčová-Medviďová, and S. Noelle. IMEX large time step finite volume methods for low Froude number shallow water flows. *Commun. Comput. Phys.*, 16(2):307–347, 2014.

[4] S. Boscarino, G. Russo, and L. Scandurra. All Mach Number Second Order Semi-implicit Scheme for the Euler Equations of Gas Dynamics. *J. Sci. Comput.*, 77(2):850–884, 2018.

[5] F. Bouchut, E. Franck, and L. Navoret. A Low Cost Semi-implicit Low-Mach Relaxation Scheme for the Full Euler Equations. *J. Sci. Comput.*, 83(1):24, 2020.

[6] S. Clain, S. Diot, and R. Loubère. A high-order finite volume method for systems of conservation laws—Multi-dimensional Optimal Order Detection (MOOD). *J. Comput. Phys.*, 230(10):4028–4050, 2011.

[7] S. Conde, S. Gottlieb, Z. J. Grant, and J. N. Shadid. Implicit and Implicit–Explicit Strong Stability Preserving Runge–Kutta Methods with High Linear Order. *J. Sci. Comput.*, 73(2-3):667–690, 2017.

[8] P. Degond and M. Tang. All speed scheme for the low Mach number limit of the isentropic Euler equations. *Commun. Comput. Phys.*, 10(1):1–31, 2011.

[9] G. Dimarco, R. Loubère, V. Michel-Dansac, and M.-H. Vignal. Second-order implicit-explicit total variation diminishing schemes for the Euler system in the low Mach regime. *J. Comput. Phys.*, 372:178–201, 2018.

[10] G. Dimarco, R. Loubère, and M.-H. Vignal. Study of a New Asymptotic Preserving Scheme for the Euler System in the Low Mach Number Limit. *SIAM J. Sci. Comput*, 39(5):A2099–A2128, 2017.

[11] S. K. Godunov. A Difference Scheme for Numerical Solution of Discontinuous Solution of Hydrodynamic Equations. *Mat. Sb., Nov. Ser.*, 47:271–306, 1959.

[12] S. Gottlieb. On high order strong stability preserving Runge-Kutta and multi step time discretizations. *J. Sci. Comput.*, 25(1-2):105–128, 2005.

[13] S. Gottlieb, D. Ketcheson, and C.-W. Shu. *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations*. WORLD SCIENTIFIC, 2011.

[14] S. Gottlieb and C.-W. Shu. Total variation diminishing Runge-Kutta schemes. *Math. Comp.*, 67(221):73–85, 1998.

[15] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43(1):89–112, 2001.

[16] H. Guillard and C. Viozat. On the behaviour of upwind schemes in the low Mach number limit. *Comput. & Fluids*, 28(1):63–86, 1999.

[17] A. Harten. On a Class of High Resolution Total-Variation-Stable Finite-Difference Schemes. *SIAM J. Numer. Anal.*, 21(1):1–23, 1984.

[18] I. Higueras, N. Happenhofer, O. Koch, and F. Kupka. Optimized strong stability preserving IMEX Runge–Kutta methods. *J. Comput. Appl. Math.*, 272:116–140, 2014.

[19] I. Higueras, D. I. Ketcheson, and T. A. Kocsis. Optimal Monotonicity-Preserving Perturbations of a Given Runge–Kutta Method. *J. Sci. Comput.*, 76(3):1337–1369, 2018.

[20] X. Y. Hu, N. A. Adams, and C.-W. Shu. Positivity-preserving method for high-order conservative schemes solving compressible Euler equations. *J. Comput. Phys.*, 242:169–180, 2013.

[21] C. A. Kennedy and M. H. Carpenter. Additive Runge–Kutta schemes for convection–diffusion–reaction equations. *Appl. Numer. Math.*, 44(1-2):139–181, 2003.

[22] S. Klainerman and A. Majda. Singular limits of quasilinear hyperbolic systems with large parameters and the incompressible limit of compressible fluids. *Comm. Pure Appl. Math.*, 34(4):481–524, 1981.

[23] R. Klein. Scale-Dependent Models for Atmospheric Flows. *Annu. Rev. Fluid. Mech.*, 42(1):249–274, 2010.

[24] R. J. LeVeque. *Numerical methods for conservation laws*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 1992.

[25] W. H. Matthaeus and M. R. Brown. Nearly incompressible magnetohydrodynamics at low Mach number. *Phys. Fluids*, 31(12):3634, 1988.

[26] G. Métivier and S. Schochet. The incompressible limit of the non-isentropic Euler equations. *Arch. Ration. Mech. Anal.*, 158(1):61–90, 2001.

[27] V. Michel-Dansac, C. Berthon, S. Clain, and F. Foucher. A well-balanced scheme for the shallow-water equations with topography or Manning friction. *J. Comput. Phys.*, 335:115–154, 2017.

[28] V. Michel-Dansac and A. Thomann. On high-precision L^∞-stable IMEX schemes for scalar hyperbolic multi-scale equations. In *Proceedings of NumHyp 2019*, SEMA SIMAI Springer Series. Springer International Publishing, 2019.

[29] S. Noelle, G. Bispen, K. R. Arun, M. Lukáčová-Medviďová, and C.-D. Munz. A weakly asymptotic preserving low Mach number scheme for the Euler equations of gas dynamics. *SIAM J. Sci. Comput.*, 36(6):B989–B1024, 2014.

[30] L. Pareschi and G. Russo. Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations. In *Recent trends in numerical analysis*, volume 3 of *Adv. Theory Comput. Math.*, pages 269–288. Nova Sci. Publ., Huntington, NY, 2001.

[31] L. Pareschi and G. Russo. Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.*, 25(1-2):129–155, 2005.

[32] P. L. Roe. Generalized formulation of TVD Lax-Wendroff schemes. *ICASE NASA Langley Research Center, Hampton, VA*, ICASE Report No 84-53, 1984.

[33] B. Schmidtmann, B. Seibold, and M. Torrilhon. Relations Between WENO3 and Third-Order Limiting in Finite Volume Methods. *J. Sci. Comput.*, 68(2):624–652, 2015.

[34] C.-W. Shu and S. Osher. Efficient implementation of essentially nonoscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439–471, 1988.

[35] J. A. Smoller and J. L. Johnson. Global solutions for an extended class of hyperbolic systems of conservation laws. *Arch. Ration. Mech. Anal.*, 32(3), 1969.

[36] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21(5):995–1011, 1984.

[37] F. Vilar, P.-H. Maire, and R. Abgrall. Cell-centered discontinuous Galerkin discretizations for two-dimensional scalar conservation laws on unstructured grids and for one-dimensional Lagrangian hydrodynamics. *Comput. & Fluids*, 46:498–504, 2011.